DELHI COLLEGE OF ENGINEERING

सत्यमेव जयते

LIBRARY

Class No.________________

Book No.________________

Accession No.________________

Borrower is requested to check the book and get the signatures on the torned pages, if any.

DELHI COLLEGE OF ENGINEERING
Kashmere Gate, Delhi

LIBRARY

DATE DUE

For each day's delay after the due date a fine of **10 P.** per Vol. shall be charged for the first week, and **50 P.** per Vol. per day for subsequent days. Text Book **Re. 1.00.**

Borrower's No.	Date Due	Borrower's No.	Date Due

ADVANCES IN ELECTRONICS

Edited by
L. MARTON
National Bureau of Standards, Washington, D. C.

Editorial Board

T. E. Allibone
H. Diamond
W. G. Dow
G. F. Metcalf
A. O. C. Nier
W. B. Nottingham
A. Rose
L. P. Smith

VOLUME I

1948

ACADEMIC PRESS INC., PUBLISHERS
NEW YORK, N. Y.

Copyright, 1948, by
ACADEMIC PRESS INC.
125 East 23rd Street
New York 10, N. Y.

All Rights Reserved

NO PART OF THIS BOOK MAY BE REPRODUCED IN ANY FORM, BY PHOTOSTAT, MICROFILM, OR ANY OTHER MEANS, WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

PRINTED IN THE UNITED STATES OF AMERICA

CONTRIBUTORS TO VOLUME I

Albert S. Eisenstein, *Department of Physics, University of Missouri, Columbia, Mo.*

Jack W. Herbstreit, *Central Radio Propagation Laboratory, National Bureau of Standards, Washington, D. C.*

R. G. E. Hutter, *Sylvania Electric Products, Inc., Bayside, N. Y.*

Mark G. Inghram, *Argonne National Laboratory, Chicago, Ill.*

M. Stanley Livingston, *Brookhaven National Laboratory, Upton, L. I., N. Y.*

Kenneth G. McKay, *Bell Telephone Laboratories, Murray Hill, N. Y.*

A. G. McNish, *Central Radio Propagation Laboratory, National Bureau of Standards, Washington, D. C.*

Kenneth A. Norton, *Central Radio Propagation Laboratory, National Bureau of Standards, Washington, D. C.*

J. A. Pierce, *Cruft Laboratory, Harvard University, Cambridge, Mass.*

A. Rose, *RCA Laboratories Division, Princeton, N. J.*

PREFACE

The last few years have seen the extremely rapid development of many branches of physical and engineering sciences. This rapid development has both its advantages and disadvantages. Its advantages being quite obvious, it seems worthwhile to discuss some of its disadvantages. With the increasing volume of knowledge in all branches, the worker in any field is unable to follow the information in a very wide field and, therefore, rapidly becomes specialized. This increasing specialization sometimes results in the adoption of a special language by publications in one branch, which may differ so much from language adopted in other branches that the research worker in a neighboring field may have to go into philological and semantic studies before being able to understand the meaning of some of the papers.

The growing number of publications creates another difficulty. It becomes more and more perplexing for the research worker to gather all information required when attacking a new subject, or when supplementing his own knowledge by information from neighboring fields. These difficulties are not new; they existed before the war though the aspect is now quantitatively different. When the total number of publications is considered it becomes imperative to produce some guide to the research worker who wishes to acquaint himself with advances in related fields.

In the field of electronics, after recognizing the need for producing summaries of some kind, we faced another difficulty. Originally, the word electronics meant that branch of science dealing with electrons. Gradually the devices using electrons were included and the time came when electronics referred to the whole branch of science and technology dealing with the devices that use free electrons. Unfortunately, the meaning of the word was not fixed even then, and, commercial publicity helping, the word electronics gradually degenerated until it not only designated the instruments using free electrons but also all the associated circuitry. By following the modern trend the editors and publishers of this volume were faced with the problem of including either a very wide selection of reviews dealing with extremely different subjects, or using a somewhat arbitrary definition of the word electronics and thus limiting the choice of materials to a reasonably restricted field.

In agreement with the publishers, the Board of Editors of these volumes preferred to choose the second alternative. The choice made is best defined by the following excerpt from the "Outline of Editorial Policy."

"The 'Advances in Electronics' is a yearly publication of articles devoted to the general field of electronics. This term 'Electronics' is often used in a rather broad sense; for the purpose of limiting the scope of this publication it seems useful to define what we wish to include here. Our intention is to publish critical and integrated reviews of specific topics in the field of physical electronics and in selected fields of engineering electronics. Physical electronics usually embraces the basic physics of charged particles (both positive and negative): emission phenomena, shaping and guiding of beams, space charge effects, interaction with matter, etc. Engineering electronics usually embraces the methods and instrumentation for practical application of such charged particles in the numerous devices which use them. We wish to impose a certain limitation on the discussion of the instrumentation: too often in technical literature a discussion of the circuitry becomes so preponderant that in some people's minds the word electronics means high frequency circuit engineering. To avoid confusion, we wish to make clear that, although the circuits are inseparably bound to the electronic devices, we desire to limit the discussion of circuits to a reasonable minimum. In thus outlining the scope of the publication, the emphasis should be on the fundamental part of the progress rather than on the accessories."

In spite of some minor difficulties (slow release of classified material, etc.), it is hoped that part of the goal has been reached in this first volume. No such work, however, can be perfect and, therefore, it is sincerely hoped that future editions may be improved by criticisms and suggestions from the readers of this publication.

To finish, I express my deepest appreciation for the generous help and advice given by the members of the Editorial Board, and also acknowledge the help of such friends who were kind enough to take over the critical reading of some of the manuscripts. And, last but not least, I am indebted to the authors whose presentations make this volume possible.

L. MARTON

Washington, D. C.

CONTENTS

CONTRIBUTORS TO VOLUME I v

PREFACE vii

Oxide Coated Cathodes

By ALBERT S. EISENSTEIN, *University of Missouri, Columbia, Mo.*

I. Introduction 1
II. Properties of the Coating 3
III. Properties of the Interface 24
IV. Properties of the Complete Cathode 29
V. Thin Oxide Film Phenomena 58
References 62

Secondary Electron Emission

By KENNETH G. MCKAY, *Bell Telephone Laboratories, Murray Hill, N. Y.*

Introduction 66
I. Pure Metals 66
II. Insulators 97
III. Composite Surfaces 114
Bibliography 120

Television Pickup Tubes and the Problem of Vision

By A. ROSE, *RCA Laboratories Division, Princeton, N. J.*

I. Introduction 131
II. Major Types of Pickup Devices 132
III. Number and Variety of Television Pickup Tubes 133
IV. Comparison of Actual and Possible Pickup Tubes 134
V. Ideal Performance 135
VI. An Experimental Realization of Ideal Performance 141
VII. Performance of Selected Pickup Devices 146
VIII. A Criterion for Noise Visibility 160
IX. Intelligence vs. Bandwidth and Signal-to-Noise Ratio . . . 163
X. Concluding Remarks 165
References 165

The Deflection of Beams of Charged Particles

By R. G. E. HUTTER, *Sylvania Electric Products, Inc., Bayside, N. Y.*

I. Introduction 167
II. Small-Angle Deflection 168
III. Large-Angle Deflection 200
References 218

Modern Mass Spectroscopy

By Mark G. Ingheam, *Argonne National Laboratory, Chicago, Ill.*

I. Introduction 219
II. General Theory 220
III. Apparatus 227
IV. Uses of the Mass Spectroscope 253
V. Commercially Available Mass Spectrometers 265
References 265

Particle Accelerators

By M. Stanley Livingston, *Brookhaven National Laboratory, Upton, L. I., New York*

I. Introduction 269
II. Direct Voltage Generators 271
III. Resonance Accelerators: The Cyclotron 271
IV. Induction Accelerators: The Betatron 278
V. Principles of Acceleration to High Energies 281
VI. The Synchrotron 294
VII. The Synchro-cyclotron 300
VIII. The Linear Accelerator 306
IX. Future Possibilities: The Proton Synchrotron 312
References 315

Ionospheric Research

By A. G. McNish, *National Bureau of Standards, Washington, D. C.*

I. Introduction 317
II. Research During World War II 320
III. Geomagnetic Effects in the F2 Layer 321
IV. Distribution of E and F1 Layers 324
V. Two Control-Point Method of Calculating Maximum Usable Frequencies 324
VI. Effects of Solar Activity 326
VII. Prediction of Ionospheric Disturbances 330
VIII. Sporadic E Reflections 332
IX. Absorption of Radio Waves 333
X. Radio Noise 338
XI. Reflections from Meteor Trails 340
XII. High-Speed Multifrequency Recorder 34
XIII. Trends of Research 34
References 34

Cosmic Radio Noise

By Jack W. Herbstreit, *National Bureau of Standards, Washington, D. C.*

I. Introduction 34
II. Jansky's Measurements 34
III. Reber's Early Measurements 349
IV. Later Measurements 354

V. The Point Source in Cygnus 356
VI. National Bureau of Standards Measurements. 356
VII. Method of Measurement. 358
VIII. Results of Measurements. 364
IX. Analysis in Terms of External Noise Factors 366
X. Field Intensities Required for Communication Services 369
XI. Effective Temperature Concept. 369
XII. Distribution of the Intensity of the Noise Sources with Direction and Frequency 370
XIII. Intensity from Small Noise Sources 375
XIV. Observed Intensity of Radio Frequency Radiation from the Sun 376
XV. Polarization of Extraterrestrial Radiation 378
XVI. Origin of Cosmic Radio Noise. 379
References 380

Propagation in the FM Broadcast Band

By Kenneth A. Norton, *National Bureau of Standards, Washington, D. C.*

I. Introduction 381
II. The Interference Due to Long Distance Ionospheric Propagation . . . 382
III. The Effects of Radio Noise on Broadcast Reception. 387
IV. The Effects of Antenna Height and Terrain on the Effective Transmission Range Over a Smooth Spherical Earth. 390
V. The Effects of Irregularities in the Terrain. 391
VI. The Systematic Effects of Terrain and of Tropospheric Ducts. 392
VII. The Tropospheric Waves Resulting from Reflection at Atmospheric Boundary Layers 402
VIII. The Combined Effects of Ducts and of Random Tropospheric Waves . . 408
IX. The Calculated Service and Interference Ranges of FM Broadcast Stations. 410
X. The Efficient Allocation of Facilities to FM Broadcast Stations. 413
XI. The Optimum Frequency for an FM Broadcast Service 414
References 421

Electronic Aids to Navigation

By J. A. Pierce, *Harvard University, Cambridge, Mass.*

I. Introduction 425
II. Prewar Methods 427
III. Wartime Developments 431
IV. Postwar Proposals. 434
V. Considerations of Range and Accuracy. 436

Author Index. 453

Subject Index 462

Oxide Coated Cathodes

ALBERT S EISENSTEIN

Department of Physics, University of Missouri, Columbia, Mo

CONTENTS

		Page
I	Introduction	1
II	Properties of the Coating	3
	1 Monolayer Theories of Emission	4
	2 The Modern Theory of Semiconductors	6
	3 Physical Structure of the Coating	9
	4 Conductivity	12
	5 Hall Coefficient	20
	6 Density of Impurity Levels	21
	7. Absorption Spectra	22
	8 Luminescence	23
III	Properties of the Interface	24
	1. Composition	24
	2 Thickness	25
	3 Conductivity	27
IV.	Properties of the Complete Cathode	29
	1 Thermionic Emission	29
	2 Photoelectric Emission	34
	3 Emission in Retarding Fields	35
	4 Emission in Accelerating Fields	37
	5 Conductivity	40
	6 Rectification	47
	7 Thermoelectric Effect	50
	8 Emission Decay Phenomena	51
	9 Sparking	55
	10 Secondary Emission	58
V.	Thin Oxide Film Phenomena	58
	1 Thickness Dependence	59
	2. Base Metal Dependence	61
	References	62

I. INTRODUCTION

In 1939 Blewett published an extensive review article[1] on "The Properties of Oxide Coated Cathodes" in which, for the first time, the literature was carefully sifted to determine whether a coherent description of the oxide cathode could be constructed from the results of more than one hundred separate and often conflicting experiments. Throughout the span of 44 years, during which oxide cathodes have been the subject of both study and controversy, experimenters have described their results in terms of existing theories. With the advent of Wilson's modern theory[2] of semiconductors in 1931 many of the oxide cathode characteristics which previously were explained on the basis of superficial "mechanical" models could then be correlated under one coherent theory. In the interim between the publication of Blewett's review and the close of World War II, a group of some thirty additional papers appeared and were listed in a literature survey[3] made by the same author in 1945.

During the war years, problems arising in the improvement and development of oxide cathodes for specific applications provided a stimulus for research into certain of the oxide cathode characteristics. As pointed out by Vick in a recent review,[4] this wartime research served 1) to investigate the thermionic emission capabilities of cathodes under microsecond pulsed conditions and 2) to study the interface region lying between the cathode base metal and the coating, the latter because of its possible influence on the cathode's thermionic emission properties. Although both of these important characteristics were recognized prior to 1941, exigencies of the war accelerated this kind of work and at the same time channeled research activities into a relatively small number of research laboratories.

It is the purpose of this review to discuss in detail some of the more recent oxide cathode researches, to re-examine certain of the earlier publications in order to ascertain whether the experimental results are consistent with the modern theory of semiconductors, and to suggest, where possible, new experiments which may lead to a more complete understanding of the oxide coated cathode.

The subject matter is grouped under three broad headings, properties of the coating, properties of the interface, and properties of the complete cathode, followed by a short discussion of thin oxide film phenomena. It is hoped that a separation of the interface and coating phenomena will lead to a better understanding of each before considering the influence of both in the complete cathode. This division, in certain cases, is an arbitrary one and future study may lead to a considerably different classification of "cause and event."

Although the thermionic emission properties of many materials have been examined, only the alkaline earth oxides combine the features of high efficiency, reasonable stability, and long life. Thoria, ThO_2, shows promise of providing even greater stability and life but at a much lower efficiency.[5] This discussion will be confined solely to the alkaline earth oxides for within this group the properties are closely related.

II. PROPERTIES OF THE COATING

Present day oxide coated cathodes are prepared by first applying carbonates of the alkaline earth elements to a base metal and then converting the carbonates to oxides in vacuum. The carbonates may consist of mixed $BaCO_3$ and $SrCO_3$ or a coprecipitated solid solution of the carbonates $(BaSr)CO_3$ although occasionally a few per cent of $CaCO_3$ is added. The base metal, usually nickel, may be an electrolytic nickel of high purity or a nickel alloy containing a few per cent of reducing elements, e.g., silicon, titanium, magnesium, manganese, etc.

While exhausting the tube containing the cathode, one applies heat to convert the carbonates to the oxides through the evolution of carbon dioxide. At the conclusion of the conversion process, sometimes referred to as "break down," the cathode is activated by heating alone (thermal activation) or through the drawing of emission current in increasing steps of current density. Preferably the activation process is carried out while the tube is under exhaust so that gases which are evolved may be completely removed. It is a common industrial practice however, to activate the cathode in the sealed off and gettered vacuum tube.

Prior to the introduction of the interface concept, oxide cathode research was directed primarily toward a study of the oxide coating, particularly the relationship between the physical and chemical properties of the coating and the thermionic emission capabilities of the cathode. Many well-intended experiments designed to examine the effect of one coating parameter have produced ambiguous results due to the difficulty of maintaining the other parameters, some possibly unknown, fixed during the experiment. Certain coating parameters are subtly interrelated, e.g., impurity content and particle size, so that unless precautions are observed, spurious effects cannot be eliminated. Changes in extraneous parameters must constantly be guarded against in experiments designed to examine the effect of only one coating variable. Humidity fluctuations, variable processing conditions, unknown base metal conditions, the formation of interfaces, etc., have undoubtedly influenced the outcome of many experiments and resulted in the existing confusion concerning the relationship of certain coating parameters to the thermionic behavior. This section deals with those coating parameters which are probably not influenced by interface considerations, yet are related to the electronic properties of the cathode in a fundamental manner. Effects of particle size, coating weight and density, and details of the conversion process are thus excluded.

1. Monolayer Theories of Emission

Prior to the introduction of the modern theory of semiconductors, cathode emission theories were based upon mechanisms for lowering the normal work function of the base metal. Due to the similarity between the activation procedure for thoriated-tungsten emitters and oxide cathodes, it was natural that similar theories were developed.

Becker[6] suggested that in active cathodes a monolayer of free barium or strontium was present on or near the external surface of the cathode and he carried out many ingenious experiments which were consistent with this point of view. Lowry[7] placed the activation monolayer at the base metal surface since his experiments indicated a decided dependence of emission on the core material. A further point of difference between the two theories, other than the exact location of the monolayer, lay in the role assigned to the cathode coating. Becker assumed the electronic conductivity of the coating to be sufficiently great that electrons could freely move from the base metal to the external surface where the total work function was located and wherein lay the "seat of emission." Lowry considered the "seat of emission" to exist at the core and the coating to be very poorly conducting so that electrons emitted at the surface of the base metal diffused through the interstices of the coating. This seemed not too improbable for the interstices of most coatings occupy nearly three-quarters of the total volume. The decay of emission with life and the low emission properties of a glazed cathode were explained as being due to a sintering of the coating which blocked the outward diffusion of electrons. A pseudo space charge of electrons in the coating interstices near the surface was used to explain the non-saturating emission characteristics. Reimann and Murgoci[8] found a correspondence between saturation of the conduction current and saturation of the emission current and, like Lowry, placed a monolayer of barium or strontium at the base metal. Transport of electrons to the outer surface of the coating was facilitated by a process of barium ion conduction through the crystals and re-emission across the interstices between crystals.

Although these theories were proposed at about the same time the Becker theory has been more generally accepted. In one experimen[9] designed to differentiate between the two models, Becker and Searst found the emission current to drop by four orders of magnitude when the coating was removed from the base metal by a mechanical shock. More recently, Jones[10] reported the removal in vacuum of only the outer surface layers of the coating and the subsequent decreases in emission by three orders of magnitude. The cathode which was deactivated by removing the entire coating could not be reactivated by thermal treatment whereas the activity of Jones' cathode was completely restored. This occurred, presumedly, due to the formation of a new monolayer on the freshly exposed surface. Certainly the external surface of the cathode cannot be neglected in theories describing electron emission from oxide cathodes though it seems incorrect to suspect that the emission is controlled only by the work function of this external surface.

The total work function is related to the energy required to carry an electron from the base metal to an infinitely distant point outside of the cathode. When atoms are adsorbed on the outer surface of a material their influence on thermionic emission is usually described by stating the change, $\Delta\phi$, which they produce in this work function, ϕ. A lowering of the potential barrier at the surface of the cathode was considered to come about due to the formation of a double layer at the surface either through a polarization of the adsorbed atoms, adatoms, through the partial ionization of the adatoms, or both. In terms of the average dipole moment per adatom, $\bar{M}$, the reduction in work function is[11]

$$\Delta\phi = 4\pi\bar{M}\eta \tag{1}$$

where η is the number of adatoms per unit area. The value of $\bar{M}$ is not entirely independent of η, particularly for large values of the concentration. The adsorption of barium atoms on the outer SrO surface, see section II-3, would probably take place through a bonding between the barium adatoms and the oxygen atoms of the exposed lattice. If the 100 plane of the SrO face centered cubic lattice lies in the surface, approximately 1.4×10^{15} oxygen sites/cm.^{2} are exposed. As judged by its interatomic spacing in the metallic lattice, barium atoms are of such a size that only about half of the oxygen sites could be occupied. Assuming that the average dipole moment remains constant, the thermionic emission should increase as η becomes larger. It is generally believed that η will not increase much beyond the value corresponding to a monolayer coverage due to the very large reduction in the oxygen bonding affinity for more than one adatom.

The electron diffraction technique has been used to examine the active emitting surface of a cathode with the expectation of confirming or denying the presence of a monolayer of adatoms. Darbyshire's[12] earlier experiments were interpreted as evidence for a surface layer barium although more recently, Huber and Wagner[13] found no conclusive evidence for the existence of this monolayer.

In view of the success already achieved by the modern theory of semiconductors, it seems possible that the "seat of emission" need not be associated with a monolayer of adatoms. The "seat of emission" may be found at the impurity centers located in the crystal but near the surface, say within an electron mean free path. The presence of lattice defects near the surface might easily give rise to a much greater concentration of impurity centers there than would be found throughout the bulk of the crystal. Until simultaneous measurements are made of the Hall coefficient and the electrical conductivity, we can only speculate concerning the electron mean free path. However, it seems not too improbable that this may be more than a few interatomic distances.

2. *The Modern Theory of Semiconductors*

Wilson's theory[2] pertaining to the impurity type of semiconductor has been notably successful in linking together a wide variety of the physical characteristics of certain of the nonmetals. Within recent years the explanation of many oxide cathode experiments has been sought on the basis of this theory. In view of the experiments of Becker and others it is assumed that a stoichiometric excess of barium, rather than oxygen or other foreign atoms, gives rise to the semiconductor characteristics. Thus we are led to the model of an *excess* impurity semiconductor, referred to as a semiconductor of the N type. A material containing an excess of the electronegative element and thus a deficiency of the electropositive element is known as a semiconductor of the P type. Few, if any, experiments are reported which show conclusively that (BaSr)O or BaO is an excess semiconductor. Nevertheless, those properties of the cathode coating which are amenable to test and have been studied agree fairly well with the behavior expected of a semiconductor of either the N or P type. Certainly the way is now open for experiments to definitely prove or disprove whether the oxide cathode actually has characteristics predicted for semiconductors, and if so to indicate whether it is of the N or P type.

Current interest in the electronic properties of the nonmetals has given rise to a group of articles[14,15,16,17] which review the basic principles of semiconductors. Following the quantum theory, the electrons in a crystal are considered to occupy certain discrete energy states or levels. Application of the exclusion principle prohibits more than two electrons from occupying the same quantum state. This may be thought of as a spreading of the normally well spaced energy levels of an isolated atom, each level giving rise to a closely spaced group of levels. Thus the allowed energy states are not uniformly distributed in energy but are grouped into bands of closely spaced levels, separated by rather wide forbidden regions. An isolated atom possesses, in addition to the normally occupied energy levels, a number of excited states representing possible energy configurations of higher total energy. Similarly, the crystal may possess one or more energy bands which are not occupied at absolute zero since the electrons prefer the states of lowest energy. These normally unfilled levels comprise the conduction band, for it is the motion of free electrons of these energies which is responsible for the electronic conductivity of the material. As the temperature is increased, some electrons may acquire sufficient energy to occupy certain levels in the conduction band. If, however, a large energy gap separates the two bands, relatively few electrons will make this transition. When the energy width of the forbidden region is large, say 10 electron volts, the material is called an insulator; if the energy gap is only a few electron volts, we have an intrinsic semiconductor. The presence of impurity atoms may add levels in this forbidden region and thus provide a source of more readily available electrons, due to the lesser energy which is required to produce a transition to the conduction band. Fig. 1 illustrates the energy level configuration of an N-type, impurity semiconductor. Possible electron energy levels are represented by the vertical scale whereas a space variation is indicated horizontally. Zero energy is taken as the electrostatic potential inside the crystal, and the electrostatic potential just outside the semiconductor is $-\Phi_v$. The energy of an electron just outside the semiconductor is $-e\Phi_v$; the energy at the bottom of the conduction band is ϵ_0, and the energy of the impurity levels is $\epsilon_0 - \Delta\epsilon$. The energy gap separating the filled band and ϵ_0 is usually large compared to $\Delta\epsilon$, hence it is not necessary to consider energy transitions from the filled band.

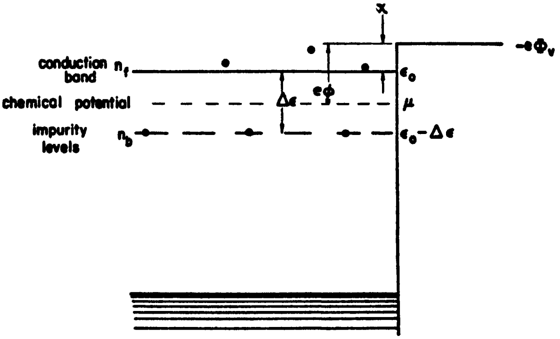


FIG. 1.—Energy level arrangement for excess impurity semiconductor.

The symbol μ represents the chemical potential of the electrons and for the N-type semiconductor, falls between ϵ_0 and $\epsilon_0 - \Delta\epsilon$.* By the chemical potential we mean the partial molar free energy of the electrons, this term occurring as a parameter in the Fermi distribution function where P, the probability that a conduction energy level ϵ is occupied, is,

---

* Conyers Herring has very recently pointed out that if the number of impurity centers exceeds the number of available electrons μ may be in the vicinity of $\epsilon_0 - \Delta\epsilon$.

$$P = \frac{1}{e^{\frac{\epsilon - \mu}{kT}} + 1} \tag{2}$$

The analogy between this and the distribution function for metals has led to the use of the term "Fermi level" for the chemical potential.

The quantity χ, representing the difference in energy between the bottom of the conduction band and a position just outside the surface, is known as the electron affinity of the crystal or the surface work function. As we shall presently see, the release of an electron from the crystal requires an energy $e\phi$; ϕ is known as the total thermionic work function. At absolute zero n_b electrons are usually considered to completely occupy the bound impurity levels whereas the conduction band is completely empty. At some higher temperature n_f *free* electrons are found in the conduction band leaving $n_b - n_f$ electrons at the impurity centers. These terms represent electron densities since a unit volume is considered in each case.

The particular physical model which the impurity levels represent is subject to considerable speculation. An N-type impurity semiconductor characterized by a stoichiometric *excess* of the metallic element may dispose the excess atoms either interstitially throughout the lattice or at normal lattice sites. This latter arrangement is achieved only by the deficiency of electronegative atoms from the same number of lattice sites as there are excess electropositive metallic atoms. Either of the two models provide a source of readily available electrons, electrons trapped in the field of the interstitial metallic ions or electrons trapped at lattice defects, caused by missing electronegative atoms. The term "F center" designates a lattice defect of the latter type. The P-type impurity semiconductor, characterized by a stoichiometric *deficiency* of the metallic atom, has impurity levels just above the filled band and conducts due to the motion of electron vacancies or "holes" in the filled band. This form of impurity semiconductor may likewise be described but using for a model the disposition of excess electronegative atoms interstitially or at regular lattice sites. In this case F centers are formed due to missing electropositive atoms. Either of these arrangements provides a "trap" for electrons which may be excited from the filled band. Although both physical models may be used for the N- and P-type semiconductors, the interstitial arrangement has been preferred for N-type materials and the F-center model is generally used to explain P-type behavior.

The semiconductor properties of oxide cathode coatings have often been attributed to interstitial barium or strontium atoms believed to contribute one or two valence electrons to the conduction band through the acquisition of sufficient energy to separate the electrons from the remaining ion. In view of atomic size considerations, use of the F-center model may be more appropriate. The Goldschmidt ionic diameters of barium and strontium are 2.86×10^{-8} cm. and 2.54×10^{-8} cm. respectively, compared with the maximum atomic interspace of about 1×10^{-8} cm. in BaO. Certainly a very distorted and considerably stressed lattice would result from the interstitial arrangement. The F-center model, wherein certain oxygen lattice sites are vacant, is capable of providing all of the electronic behavior required of an interstitial impurity center without markedly distorting the lattice.

Certain details concerning the nature of the filled band of the alkaline earth oxides have been determined. O'Bryan and Skinner[18] have made use of the line breadth of long wavelength x-ray emission spectral lines to evaluate the widths of the uppermost filled bands. They are: BaO, $8.4 eV$; SrO, $9.2 eV$; and CaO, $10.8 eV$. These uppermost filled bands of the oxides are believed to originate from the L shell, *p* electrons of the oxygen ion. If the alkaline earth oxides behave as N-type semiconductors, details of the filled band are of little consequence in determining the electronic properties of the coating at normal operating temperatures although this information is of possible use in interpreting the optical absorption spectra.

3. *Physical Structure of the Coating*

X-ray diffraction techniques have provided a valuable tool for examining the physical structure of the oxide cathode coating. Vegard's rule is found to apply to solid solutions of the oxides, in that the lattice constant of the solid solution varies linearly[13,19] with composition and continuously from 100% BaO to 100% SrO (see Fig. 2). Burgers[19] showed that when a mechanical mixture of $BaCO_3$ and $SrCO_3$ was decomposed in vacuum a simple conversion to the mixed oxides, BaO and SrO, occurred. Further heating caused the formation of a true solid solution, (BaSr)O. In a similar investigation, Benjamin and Rooksby[20] reported an increase in the d.-c. emission from the homogeneous oxide solid solution over the emission obtained from the mixed oxides. This dependence was related to the well known variation of d.-c. thermionic emission with solid solution composition, first observed by Benjamin and Rooksby,[21] Fig. 3. A similar dependence of pulsed emission on solid solution composition was recently reported by Veenemans[22] and Widell.[23] Considering that the oxide coating behaves as an excess impurity semiconductor, a qualitative explanation[13] was offered for this behavior. The lattice of the oxide solid solution is distorted owing to the different size metallic atoms occupying a common lattice. As a result, the energy $\Delta\epsilon$ required to release an electron from an impurity

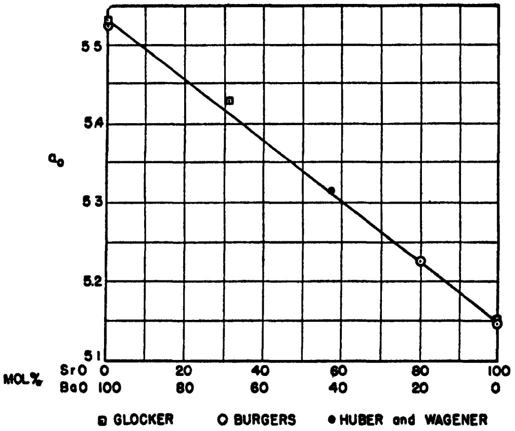


Fig. 2.—Variation of lattice constant with composition for solid solutions of (BaSr)O. See ref. 13.

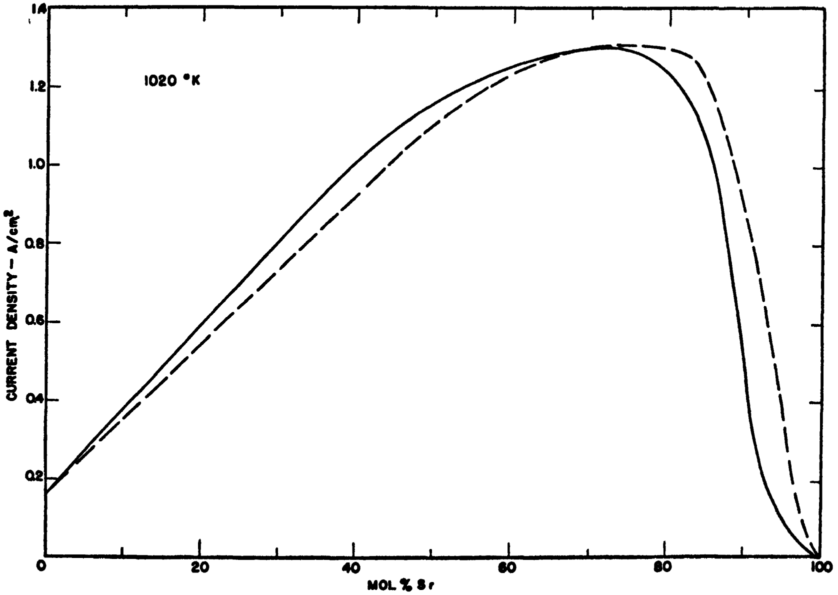


Fig. 3.—Dependence of d.-c. emission on mol per cent Sr content of (BaSr)O cathode. Solid line, carbonate content before heat treatment; broken line, oxide content after activation. See ref. 21.

center to the conduction band may be less than that required if the impurity had been in the more perfect lattice of the pure oxides. As we shall see later, conductivity and thermionic emission depend on the number of free electrons in the conduction band. Values of $\Delta\epsilon$ have been determined only for SrO and (BaSr)O (see section IV-5), and these are in agreement with the above hypothesis. Thus it seems that the position of the energy levels of the impurity atoms relative to the conduction band is subject to change, depending on the neighborhood surrounding the impurity center.

Changes which occur in the oxide composition are due principally to the preferential evaporation of Ba or BaO from the external surface of the cathode. The lower vapor pressure SrO remains and within a short time, usually by the completion of conversion, the outer surface is free of BaO. The formation of a BaO-deficient surface must result from the fact that the rate of surface evaporation exceeds the rate of migration to the surface. Gaertner,[24] Darbyshire,[25] and Huber and Wagener,[13] using electron diffraction techniques, agree that in activated, initially equal molar cathodes the surface consists of nearly pure SrO. Burgers[19] showed that relatively thick BaO-deficient surface layers could be detected using x-rays. A method of analysis[26] was devised which allows a calculation of the variation of composition with depth below the surface to be made from the x-ray scattering pattern. Fig. 4 shows the results of one such analysis. In 115 hours a surface layer containing less than 10% BaO covers the surface to a depth of about 10^{-5} cm. while at 1205 hours this layer has a thickness of 3×10^{-4} cm. Changes in the

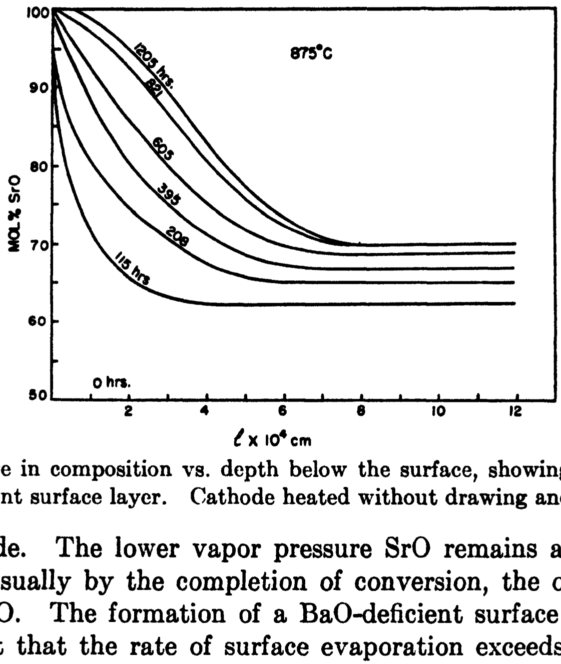


FIG. 4.—Change in composition vs. depth below the surface, showing formation of BaO-deficient surface layer. Cathode heated without drawing anode current.

bulk composition are seen by comparing the respective compositions at a depth of say 10^{-8} cm. Although reducing elements present in the base metal strongly influenced the bulk loss of BaO, impoverishment of the surface layers was influenced to a lesser extent.

The influence of the nearly pure SrO surface layers on the oxide cathode's emission properties remains somewhat problematical. Since the physical nature of the external surface is relatively independent of the bulk oxide composition, it seems not unreasonable that the electron affinity χ should be found independent of composition also (see section IV-5). Under pulsed emission conditions, the lower conductivity of the SrO outer layer may influence[27] the maximum current which may be taken. Undoubtedly the end of cathode life is reached at about the same time the concentration of BaO is depleted in the entire coating. However, it is not certain that the two are related only by the functional variation shown in Fig. 3. A reduction of the BaO content over a long period of time may influence the interface properties so as to hasten the end of life.

The technique of electron projection tube microphotography has been used extensively to examine the surface of an emitting oxide coating. A pattern of light and dark patches are observed on the viewing screen, and it is generally assumed that each emitting patch is due to the emission from a single oxide particle, although these patches might represent interstices of the coating. Heinze and Wagner[28] observed the emission pattern of a small group of well separated oxide crystals on a base metal and found it to be identical with an optical photograph showing the location of the crystals. If the particles were indeed single crystals, the emission must have originated at the crystal faces. Mecklenburg[29] used a projection tube capable of magnifications up to 50,000× which permitted the identification of emission patches as small as 400 A. in size. Such patches may well represent emission from single crystals since the size is in fair agreement with the lower limit of the oxide crystal size as determined by x-ray line breadth measurements.[30]

A working model of the normal oxide coating will consist then of a relatively thick layer of (BaSr)O which changes in composition near the surface to pure SrO.

4. Conductivity

Measurements of the electronic conductivity of materials and the comparison of these results with theory has provided an effective means for classifying many substances. The specific electronic conductivity of a sample σ, defined as the current-voltage ratio between opposite faces for a cube of material of unit volume, is related to the number of conduction electrons n_f and their mean mobility v. Thus,

$$\sigma = n_f e v \tag{3}$$

where e is the electronic charge.
Now, relating[15] the mean mobility to the electron mean free path l which is assumed to have a constant value l_0, at a temperature T,

$$\sigma = \frac{4 n_f l_0 e^2}{3(2\pi m^* kT)^{\frac{1}{2}}} \tag{4}$$

where k is Boltzman's constant. m^* is known as the effective mass of the electron being in general unequal to the electronic mass for electrons in a filled band but replaceable by the electronic mass for electrons in the conduction band. The use of this term does not imply a true change in mass but rather a change in the behavior of the electron in an applied external field.

The density of electrons in the conduction band is related to the density of impurity levels and to a Boltzman factor. This relationship[31] is,

$$n_f = \sqrt{2}\, n_b^{\frac{1}{2}} \frac{(2\pi m^* kT)^{\frac{3}{4}}}{h^{\frac{3}{2}}} e^{-\frac{\Delta\epsilon}{2kT}} \tag{5}$$

where h is Planck's constant.
Introducing this value in eq. (4) gives,

$$\sigma = \frac{4\sqrt{2}}{3} n_b^{\frac{1}{2}} \frac{e^2 l_0}{h^{\frac{3}{2}}} (2\pi m^* kT)_{\frac{1}{4}} e^{-\frac{\Delta\epsilon}{2kT}} \tag{6}$$

Except for very high temperatures, $\Delta\epsilon/2 \gg kT$, and the temperature variation of conductivity is due primarily to the exponential term. Eq. (6) is usually related to the experimentally observed temperature variation of conductivity which is written,

$$\sigma = Ke^{-\frac{Q}{kT}} \tag{7}$$

where $2Q$ is equivalent to the thermal activation energy $\Delta\epsilon$. The coefficient K which determines the magnitude of the conductivity may, over a limited range of temperature, be regarded as a function only of the density of activation centers n_b.

Making use of the relationship[31] for the location of the chemical potential μ,

$$\mu = \epsilon_0 - \frac{\Delta\epsilon}{2} + \frac{kT}{2} \ln\left[\frac{n_b h^3}{2(2\pi m^* kT)^{\frac{3}{2}}}\right] \tag{8}$$

and substituting for $\Delta\epsilon/2$ in eq. (5), we may express the density of conduction electrons n_f,

$$n_f = \frac{2}{h^3}(2\pi m^* kT)^{\frac{3}{2}} e^{\frac{\mu - \epsilon_0}{kT}} \tag{9}$$

and substituting this value in eq. (4),

$$\sigma = \frac{8l_0e^2}{3h^3}(2\pi m^* kT) e^{\frac{\mu - \epsilon_0}{kT}} \tag{10}$$

This conductivity expression contains the density of activation centers n_b implicitly in the chemical potential μ. An increase in n_b, say during activation of the cathode, will bring about an increase in μ and thus increase the conductivity of the coating. The presence of n_b contained implicitly in the exponential term of the latter conductivity expression does not invalidate the earlier comparison of the experimental and theoretical temperature dependence. A plot of the logarithm of specific conductivity σ, as a function of the reciprocal temperature $1/T$ over a limited range, should, if the sample behaves as a semiconductor, yield a linear variation. The slope of this line is $\Delta\epsilon/2k$ from which $\Delta\epsilon$ may be evaluated. Unfortunately, a test of this type is a necessary but not a sufficient condition for electronic conductivity. Ionic conductivity also obeys an exponential temperature variation, hence measurements of transport numbers are required to fix the nature of the charge carrier.

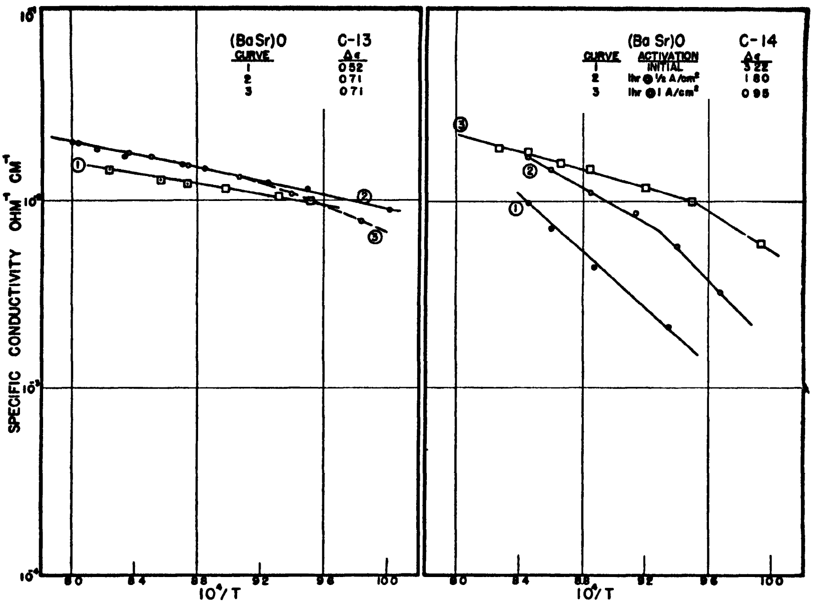
The importance of the conductivity of oxide cathode coatings has led to a number of investigations, carried out usually under differing conditions so that direct comparisons of the results are difficult. Conductivity measurements are frequently made on bulk samples of the oxide. However, when a comparison with other of the cathodes' properties is desired, the conductivity of emitting cathode coatings may be determined. Measurements of the latter type are frequently influenced by the presence of the coating-base metal interface, hence the discussion of these studies will be deferred until a later section dealing with the properties of the complete cathode. The true specific conductivity of coating materials has not been measured, but rather an effective specific conductivity which is a function of the particle size, density of packing and degree of sintering.

Blewett's[1] survey of the conductivity studies prior to 1939 gave values of the specific conductivity of BaO and (BaSr)O at 1000°K. varying from 10^{-6} ohm^{-1} cm.$^{-1}$ to 10^{-2} ohm^{-1} cm.$^{-1}$ Conductivities less than 10^{-4} ohm^{-1} cm.$^{-1}$ were reported for unactivated or loosely packed samples whereas all activated oxide coatings gave conductivities above 5×10^{-3} ohm^{-1} cm.$^{-1}$ Meyer and Schmidt[32] compressed a layer of BaO between nickel blocks and determined the conductivity using an a.-c. method.

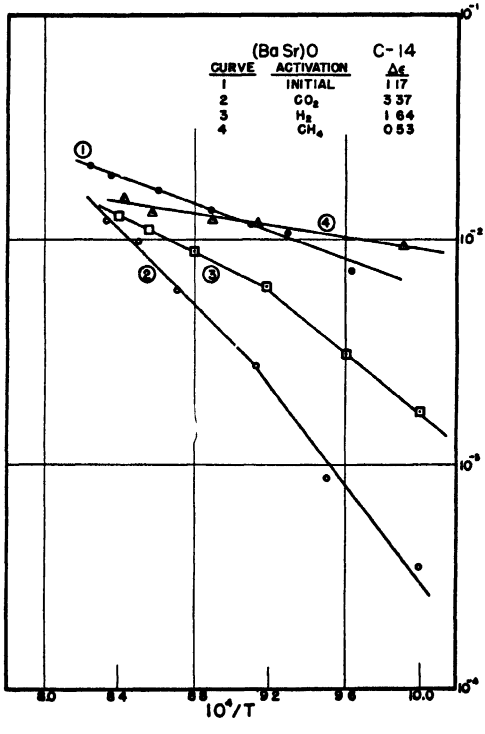
The use of nickel contacts raises the possibility that an interface conductivity is included in their experiments which otherwise represent an excellent treatment of the problem. Thermal activation of their sample was found to increase the conductivity at 1000°K. by a factor of 500. The activated state of BaO showed a temperature behavior described by eq. (7) with a thermal activation energy $\Delta\epsilon$ of $1.2eV$. Unactivated BaO showed a much higher activation energy, $\Delta\epsilon$ for this state being about $5eV$ and indicating perhaps, the behavior of an intrinsic semiconductor in which the electrons were supplied from the filled band, $5eV$ below the bottom of the conduction band. A plot of log σ vs. $1/T$ for this unactivated state did not yield a single straight line in accord with eq. (7) but instead, two lines intersecting at a point corresponding to about 1200°K. The low temperature slope corresponded to a somewhat higher activation energy than did the high temperature section.

An experimental arrangement which avoids the objection of including the interface conductivity in measurements of the coating conductivity is described by Eisenstein.[33] The oxide sample sprayed onto the surface of a MgO ceramic rod contains two imbedded platinum contactors and between these, two platinum probes. A current flowing between the contactors gives rise to a potential drop in the oxide between the probes which is measured using a null current method. If an interface is formed on the platinum probes, it will not enter into the coating conductivity measurements as zero current is taken through this interface. A knowledge of the current passing between the contactors, the potential developed between probes, the oxide thickness, and the probe separation was used to evaluate the effective specific conductivity of the sample.

Curves 1, 2, and 3 of Fig. 5(A) represent the temperature variation of conductivity of a well activated sample, C-13. A thermal activation energy of only 0.52 to 0.71 eV is indicated from the slopes of these curves. Activation of sample C-14 by the passage of a high current density is shown in Fig. 5(B) in which a change in activation energy from 3.2 to $0.95eV$ occurs. In Fig. 5(C), the same sample was deactivated and then reactivated by chemical reactions with gases introduced into the vacuum system. Measurements of the conductivity were made in vacuum immediately following the treatment in gas at about 1100°K. Assuming that a stoichiometric excess of barium enhances the conductivity, these curves may be explained as follows. Deactivation from curve 1 to curve 2 results from a decrease in the amount of excess barium n_b of eq. (6), through oxidation by carbon dioxide. The introduction of hydrogen or CH_4 causes a reduction of the oxide to increase the excess barium content and thus enhance the conductivity. Thus we see that chemical reactions are equally as effective as thermal activation and large conduc-


SPECIFIC CONDUCTIVITY OHM⁻¹ CM⁻¹
10⁻¹
10⁻²
10⁻³
10⁻⁴
(BaSr)O C-13
CURVE Δε
1 0 52
2 0 71
3 0 71
①
②
③
80 84 88 92 96 100
10⁴/T
(Ba Sr)O C-14
CURVE ACTIVATION Δε
1 INITIAL 3 22
2 1hr @ ½ A/cm² 1 80
3 1hr @ 1 A/cm² 0 95
①
②
③
80 84 88 92 96 100
10⁴/T

(A) (B)


(Ba Sr)O C-14
CURVE ACTIVATION Δε
1 INITIAL 1 17
2 CO₂ 3 37
3 H₂ 1 64
4 CH₄ 0 53
①
②
③
④
10⁻¹
10⁻²
10⁻³
10⁻⁴
8.0 84 88 92 96 10.0
10⁴/T

tion currents for creating a high state of coating activation. It is well known[34,35] that reducing materials present in the base metal of a cathode will, under certain conditions, give rise to an enhanced thermionic emission. This may be due to some extent to an increase in the coating conductivity.

Although there is general agreement between the experimental conductivity results and the theoretical predictions for any particular state of activation, certain details of this behavior require further comment.

The change of slope which occurs with changes in the states of activation, Fig. 5(C), is not predicted on the basis of the previous theoretical discussion, but is frequently observed in semiconductors, e.g., Al_2O_3, TiO_2, Fe_2O_3, and ZnO. In addition, some of the curves show a "break" similar to that found by Meyer and Schmidt.[32] Nijboer[36] has suggested that the number of impurity levels n_b may exceed the number of available electrons n_e, particularly in the case of semiconductors in which the impurity consists of excess atoms of one constituent of the crystal. Having thus provided additional vacant impurity levels $(n_b - n_e)$ for electrons making downward transitions from the conduction band, if $n_f \ll (n_b - n_e)$, then

$$n_f = 2 \frac{n_e}{n_b - n_e} \frac{(2\pi m^* kT)^{\frac{3}{2}}}{h^3} e^{-\frac{\Delta\epsilon}{kT}} \tag{11}$$

However, if $n_f \gg (n_b - n_e)$, the effect of additional vacant levels is negligible and the previously considered relationship, eq. (5), is valid if we replace n_b by n_e. Thus, at low temperatures the slope of a ln σ vs. $1/T$ plot should correspond to $\Delta\epsilon/k$ whereas at a somewhat higher temperature the slope is $\Delta\epsilon/2k$. The critical temperature T_c in the neighborhood of which the slope changes, is given when $n_f = (n_b - n_e)$, then,

$$\frac{(n_b - n_e)^2}{2n_e - n_b} = \frac{(2\pi m^* kT_c)^{\frac{3}{2}}}{h^3} e^{-\frac{\Delta\epsilon}{kT_c}} \tag{12}$$

For a given state of activation, this theoretical argument is capable of explaining a change in slope by a factor of 2, a value not entirely consistent with that observed in the previous experiments. Nor is this argument particularly attractive for explaining the variation in slope with the state of activation, for it would be necessary to postulate a change in the number of "frozen in," vacant impurity levels due to an exposure of the sample to gases. This does not seem to be a likely occurrence.

The appearance of the product $n_f l_0$ in eq. (4) prevents a unique determination of either of these fundamental parameters from a measurement of the conductivity alone. Supplying a probable value of l_0, as indicated by the mean free path in other semiconductors, leads to at least an order of magnitude for n_f and n_b. Eq. (4) may be written,

$$n_f = 2.6 \times 10^9 \frac{\sigma T^{\frac{1}{2}}}{l_0} \tag{13}$$

for σ in ohms^{-1} cm.$^{-1}$ Making use of the experimental value of σ as 10^{-2} ohm^{-1} cm.$^{-1}$ at 1000°K., for probable values of l_0 between 2×10^{-8} cm. and 10^{-7} cm. and $\Delta\epsilon = 1eV$, corresponding values of n_f and n_b are shown in Table I.

TABLE I. Values of n_f and n_b computed from eq. (13) and eq. (5), respectively, using $\sigma = 10^{-2}$ ohm^{-1} cm.$^{-1}$, $T = 1000$°K., $\Delta\epsilon = 1.0eV$, $m^* = m$.

l_0 (cm.)	n_f/cm.^{3}	n_b/cm.^{3}
2×10^{-8}	4.1×10^{16}	6.6×10^{17}
10^{-7}	$8\ 3 \times 10^{15}$	4.9×10^{16}

The use of larger values of l_0 or smaller values of $\Delta\epsilon$ results in a breakdown of eq. (5), which is applicable only when $n_f \ll n_b$. When the number of conduction electrons approaches the number of impurity levels a reduction in the value of $\epsilon_0 - \mu$ should occur, giving a conductivity slope less than $\Delta\epsilon/2k$. Under these conditions, the slope should vary slowly with temperature but not in a manner to explain the "breaks" in the curves. Whether the change of slope with activation, Fig. 5(C), can be explained on this basis is problematical and must await further conductivity measurements covering a wide range of temperatures.

Two entirely qualitative theories have been proposed to explain the decrease of slope, $\Delta\epsilon/2k$, with increasing impurity concentration n_b. Mott[36] suggests that as the concentration of impurity centers increases the average distance between such centers in the crystal decreases bringing about a mutual interaction of the bound electrons. This may cause a reduction in the binding energy and the value of $\Delta\epsilon/2k$. It is not necessary that the impurity centers become nearest neighbors for this interaction to occur as the bound electrons may exist in orbits of radii several interatomic distances from the interstitial ion or vacant lattice site. Burton[37] proposes a similarly satisfying qualitative explanation based on the presence of a few electron traps, e.g., vacant surface lattice sites, in the forbidden energy region between the filled band and the impurity levels, Fig. 1. Following the interpretation of μ suggested in section II-2, in the absence of impurity atoms the chemical potential would be located approximately midway between ϵ_0 and the top of the filled band. For small values of n_b, much less than the number of electron traps, the bound electrons would shortly find themselves occupying the lower energy electron traps and the n_b impurity levels would be nearly vacant. In this case μ would represent an energy level less than $\epsilon_0 - \Delta\epsilon$ and the slope of the conductivity plot, $(\mu - \epsilon_0)/k$, eq. (10), would be considerably greater than $\Delta\epsilon/2k$. As n_b increases, the electron traps will become completely filled with a corresponding reduction of $(\mu - \epsilon_0)/k$ until the limiting value $\Delta\epsilon/2k$ is reached.

Throughout this discussion it was assumed that the electron mean free path is independent of the state of activation and temperature. In general, the concentration of impurity centers is so small that interstitial atoms or vacant lattice sites will have a negligible effect on the mean free path. Over a wide temperature range, 600–1400°K., such as that covered in the various conductivity studies, the mean free path cannot be regarded as a constant. For this reason, measurements of the Hall coefficient together with conductivity determinations are required.

If the activation of oxide cathode coatings is due to an excess of barium either as an interstitial ion or at a lattice position, two valence electrons should be available for excitation to the conduction band. The removal of the first of these electrons may possibly be accomplished with less energy than is required to remove the second. Such a model postulates two energy levels and two slopes in the conductivity curve, the high temperature slope having a larger value than that below the "break" in the curve. That this has not been observed experimentally may be due to its occurrence in a temperature range above or below that covered in the previous experiments. Similarly, all semiconductors should show a "break" in the conductivity slope at high temperatures when transitions from the filled band become appreciable.

Many of the nonmetals exhibit an enhanced conductivity on exposure to light. Photoconductivity[38] thus induced may be divided into two components; the application of an electric field gives rise to a primary current which ceases with the removal of the illumination and to a secondary current which continues for some time following the illumination. This primary current results from an increase in n_f due to photon excited transitions from the filled band or from the impurity levels depending upon the wavelength of the exciting radiation. The secondary current is due to a trapping of electrons at energy states near the conduction levels so that a small additional thermal excitation is required for their release to the conduction band.

Photoconductivity in oxide coated cathodes is not reported in the literature and it is doubtful if measurements of its occurrence have been attempted. This is a phenomena whose investigation and study would add to our knowledge of possible energy transitions as well as trapping states of energy near the conduction levels.

5. *Hall Coefficient*

The usefulness of Hall effect measurements on semiconductor materials was indicated in sections II-2 and II-4. A single determination of the Hall coefficient at one temperature should define the material as being either an N- or P-type semiconductor and give the density of conduction electrons n_f or holes n_h. Measurements of this coefficient as a function of temperature, when combined with conductivity values over the same temperature range, allow a determination of the electron mean free path and its temperature dependence. As yet, measurements of the Hall coefficient for oxide cathode coatings are not reported in the literature, hence this represents a fertile field for future experimental studies.

This effect can best be determined using a strip of material (compressed powder or a single crystal) along which a current is passed. Probe electrodes are attached to opposite sides across the strip and may be adjusted in relative position to locate an equipotential plane although this is not absolutely necessary. The application of a magnetic field normal to the strip and at right angles to both the direction of current flow and to a line connecting the probes produces a displacement of the equipotential plane. A potential difference now appears between the probes and is known as the Hall emf, E_H. This is related[39] to the magnetic field intensity H, the current I, the strip thickness t, and a constant of the material R.

$$E_H = R\frac{HI}{t} \tag{14}$$

R is known as the Hall coefficient and for some semiconductors its value is in the range of 10^{-7} volts-cm./ampere-gauss at a temperature at which the conductivity is easily measured. Thus the use of large magnetic fields and high current densities are required to produce a measurable emf.

Making use of the usually low value of n_f, the density of electrons in the conduction band, these electrons are assumed to obey the Maxwell-Boltzman statistics which leads to an evaluation[39] of R; in electrostatic units,

$$R = -\frac{3\pi}{8n_f ec} \tag{15}$$

where c is the velocity of light. In the case of hole conduction n_f is replaced by n_h with a corresponding change in sign. The temperature variation of R is due entirely to the dependence on n_f whose temperature variation is given in eq. (5). The sign of the Hall coefficient cannot always be used to distinguish between ionic and electronic conductivity for the Hall emf resulting from ionic conduction is too small to allow its detection.

6. Density of Impurity Levels

Although a reasonable experimental verification of the temperature dependence of σ, eq. (6), has been obtained, the dependence of σ on n_b is by no means certain. As indicated in Table I the probable value of n_b is only about 10^{17} or .01% of the total atoms present. Microchemical techniques have been used to detect the presence of a stoichiometric excess of metallic atoms in activated oxide coatings. A water vapor reaction was used by Bredennikowa,[40]

$$\mathrm{Ba} + \mathrm{H_2O} \leftrightarrows \mathrm{BaO} + \mathrm{H_2}$$

in which the quantity of hydrogen produced was used to determine the excess barium present. This method was criticized[41] as not being selective in its action on only barium in the cathode. In the experiments of Prescott and Morrison[41] the reaction,

$$\mathrm{BaO} + \mathrm{CH_4} \leftrightarrows \mathrm{Ba} + \mathrm{CO} + 2\mathrm{H_2}$$

carried out at 1200°K. for 15 minutes was used to produce excess barium with a subsequent increase in thermionic emission. The close relationship between conductivity and emission will be discussed in a later section. A second chemical reaction was then employed to determine the amount of excess barium liberated during the activation treatment.
This was,

$$\mathrm{Ba} + \mathrm{CO_2} \leftrightarrows \mathrm{BaO} + \mathrm{CO}$$

carried out at 1200°K. for 2 hours, from which a quantitative gas analysis for carbon monoxide was related to the amount of barium entering into the reaction.

A correlation of the d.-c. thermionic emission and the excess barium content was made by Prescott and Morrison. Over the range from 1×10^{18} to 5×10^{18}/cm.^{3} little change in activity was observed, but below 10^{18}/cm.^{3} a marked decrease in emission occurred.

More recent work on the excess barium content of oxide cathodes has been reported by Wooten[42,43] who applied the hydrogen evolution method to active cathodes which had been removed to a separate reaction chamber in vacuum. The amount of excess barium actually present in an active oxide cathode was much less than the total free metal present in the original tube, most of the excess barium having evaporated to the walls of the tube. A careful analysis of sources of error in this type of experiment led to the belief that the actual coating content of excess barium was of the order of $10^{18}/\text{cm.}^3$.

The use of tagged, radioactive barium atoms has been suggested as a possible means for measuring excess barium densities less than that conveniently handled using microchemical techniques. An experiment of this nature might easily yield values for the diffusion rates of barium through lattices of BaO, SrO, and (BaSr)O, quantities which are desirable for future discussions of the effect of oxide composition on thermionic emission.

7. *Absorption Spectra*

Measurements of the absorption spectra of the nonmetals are *not* always useful in confirming details of the energy level configuration.[44] The removal of an electron from an impurity center by optical excitation may leave the surrounding lattice in an excited state of energy U. A series limit in the absorption spectra is related to the optical activation energy ΔE by,

$$\lambda_L = \frac{hc}{\Delta E} \tag{16}$$

where λ_L is the wavelength at which the series limit occurs. Lattice vibrations broaden the absorption lines producing an absorption band at relatively high temperatures. In general, the optical activation energy is greater than the thermal activation energy since,

$$\Delta E = \Delta\epsilon + U \tag{17}$$

Only if U is known can we accurately predict the location of the absorption band. For many materials $U \leq \Delta\epsilon/2$ and an assumption of this sort for BaO leads to a value of ΔE which places the location of an absorption band in the near infrared region of the spectrum. Although the position of this absorption band is not particularly useful, the shape of the absorption band and the value of the absorption coefficient are useful in computing[45] the density of absorbing centers, n_b.

Excitation of electrons from the filled band to the conduction band should produce absorption in the near ultraviolet region of the spectrum. If we assume the width of the forbidden region to be $5eV$, as indicated from conductivity studies, this absorption band will appear at about 2500 A.

In the far ultraviolet, true characteristic absorption by all atoms of a crystal takes place giving rise to a very large absorption coefficient. The first band of this absorption is associated with the removal of an electron from the electronegative ion and represents an energy somewhat greater than the combined widths of the uppermost filled band, the forbidden region and the conduction band. For BaO this energy would be at least 14eV, placing the band at less than 1000 A.

As yet, no systematic studies are reported which attempt to locate and make use of the absorption bands of the alkaline earth oxides. All of these oxides are colorless which indicates the absence of optical activation energies between 1.8 and 3.1 eV, values which are consistent with the proposed energy level configuration.

8. Luminescence

Fluorescence of oxide cathode coatings has undoubtedly been observed by many experimenters who perchance have bombarded a "cold" cathode with electrons or ions. In the visible region of the spectrum this fluorescence appears blue-green.

Luminescence is a term which includes both the fluorescence radiation which arises due to excitation by electrons, ions, ultraviolet, etc., and the after-glow, known as phosphorescence, which continues after the excitation is removed. Both fluorescence and phosphorescence in semiconductors have been associated with the presence of impurity levels. The role of the impurity center is not always clear although it usually provides a place where an excited electron can recombine with a hole or with the impurity center itself and radiate its excess energy. The emission spectra and hence the color of the fluorescence radiation is determined by the kind of impurity center which is involved. At room temperature this fluorescence appears over a broad band of wavelengths, but as the temperature is lowered the band becomes narrow and at low temperatures it may become a single line. Phosphorescence is attributed to the trapping of electrons in the conduction band at some distance from an impurity center with which it must recombine before emission can occur. Two types of phosphorescence emission decay are known; one in which the intensity decreases exponentially, $e^{-\alpha t}$, and one in which the decay is $t^{-\beta}$, where t represents the time. and α and β are constants. For the group of luminescent materials developed for use as cathode ray screens, there is indirect evidence that the exponential decay is associated with the substitutional type of activation center whereas the interstitial type of impurity center gives rise to the $t^{-\beta}$ decay.

Ewles[46] has examined the luminescent spectra of CaO, SrO, and BaO excited by electron bombardment at a temperature of liquid air. The emission spectra were found to be characteristic of the parent lattice and hence unchanged due to the presence of activators. Changes in the relative intensities of the sets of bands were attributed to the effect of impurities. Four broad lines appeared in the ultraviolet between 3780 and 3970 A. for SrO. In the visible region three bands were observed, at 4600, 5600, and 6450 A. BaO gave three unresolved peaks, at 4650, 5620, and 5970 A. These results seem to indicate that a direct role cannot be assigned to the impurity atoms. No attempt was made to observe the conductivity or the thermionic activity of the specimen, hence it is not certain that these samples corresponded to active states of the oxide. Here again is an untouched field for research, one which might supply valuable information concerning possible kinds of activators for the alkaline earth oxides as well as an indication of the physical disposition of these activators.

III. Properties of the Interface

The interface region lying between the base metal and the coating is now recognized as an important cathode parameter. The specific influence of the interface on the thermionic emission properties of the cathode is not always completely clear, for undoubtedly the type of interface and the conditions under which the emission is taken allow the role of the interface to change.

In this section we shall be concerned with the properties of an interface layer which differs in chemical constitution from the oxide coating. These compounds arise due to solid state chemical reactions which take place between the oxide coating and certain constitutents of the base metal.

Two additional forms of interfaces should be considered, for in the absence of an interface compound either one might provide a similar electronic behavior. The absence of activation centers in the layers of coating adjacent to the base metal or a vacuum layer caused by a poor mechanical bonding of the coating may provide a low conductivity, series element to the flow of current and thus influence the cathode's behavior.

1. Composition

Some years ago, Arnold[47] reported the presence of the compound $BaPtO_3$ at the interface of cathodes prepared on a platinum base metal, although the method used in its identification was not clear. Use of the technique of x-ray diffraction analysis for interface identification was first reported by Rooksby[48,49] who found barium aluminate, $BaAl_2O_4$, at the interface for a base metal of nickel alloyed with 2% of aluminum. Subsequent studies led to a report[50] of barium orthosilicate, Ba_2SiO_4, for a 0.4% silicon-nickel alloy base and barium orthotitanate, Ba_2TiO_4, for a 0.23% titanium-nickel alloy base. Cathodes prepared on a magnesium-nickel alloy were found[51] to have an interface layer containing magnesium oxide, MgO.

Eisenstein and Fineman[52] presented evidence for the existence of three possible interface compounds in cathodes having a chromium-nickel alloy base or a chromium plated nickel base. Two of these compounds were unstable and reverted to the third on exposure to air. Although the stable form exhibited a close packed hexagonal structure, none of the compounds were identified. Consistent but weak diffraction patterns were observed from the interface on a pure nickel base but too few lines were obtained to make identification possible.

It is interesting to note that when both barium and strontium are present in the oxide, only barium enters into the interface compound. Use of a pure SrO coating gives rise to the interface compounds Sr_2SiO_4, Sr_2TiO_4 and the stable strontium equivalent of the chromium containing compound. All are isomorphic with the equivalent barium compounds which have slightly larger lattice constants. The solid state chemical reaction between $BaCO_3$ and SiO_2 (powdered quartz) has been studied extensively[53] and although four silicates of barium are known[54] only the orthosilicate is found at the oxide cathode interface. At temperatures above 1050°K. the orthosilicate rather than the metasilicate, $BaSiO_3$, is formed from equal molar mixtures of $BaCO_3$ and SiO_2. This led to an incorrect first report[52] that the silicon interface was $BaSiO_3$; this was corrected in a later paper.[55] Similar reactions[33] were found to occur between $BaCO_3$ and silicon, for it is probably this reaction which produces the interface during the conversion process.

The structure of Ba_2SiO_4 is reported[56] as being similar to that of K_2SO_4 with the space group D_{2h}^{16} and lattice constants of 5.7, 10.1, and 7.5 A. for *a*, *b* and *c* respectively. Solid solutions of $(BaSr)_2SiO_4$ may be prepared although these are not found at the interface. Rooksby believes the structure of Ba_2SiO_4 and Ba_2TiO_4 to be isomorphous.

2. Thickness

The interface, appearing as a series element to the flow of electrons from the base metal into the coating, may under certain conditions limit the emission current from the cathode. As pointed out by Herring,[57] this will occur only if an appreciable fraction of the voltage impressed across a diode actually appears across the interface layer. This might at first seem to be a trivial case but experiments which we shall consider later show that it may occur. The electrical conductivity of the interface and the interface thickness determine the voltage which appears across this layer. Application of a barrier layer theory to the transport of electrons from the base metal into the coating likewise requires a knowledge of the interface conductivity and thickness.

When the coating is removed from an oxide coated cathode, the presence of an interface layer is frequently confirmed by its color. The silicate, titanate, and aluminate interfaces are gray; the chromium interfaces are green, brown, or black and interfaces on molybdenum are blue or red. Ba_2SiO_4 is normally colorless, hence its gray appearance at the interface may be due to interspersed, finely divided nickel particles formed by evaporation into the porous coating. Fig. 6 shows a group of cathode samples. On the left is an uncoated sleeve of a silicon-nickel alloy adjacent to which is a similar carbonate coated sleeve. On the right are three Ba_2SiO_4 interfaces formed on converting the carbonates by heating each for a fixed time and then removing the oxide.

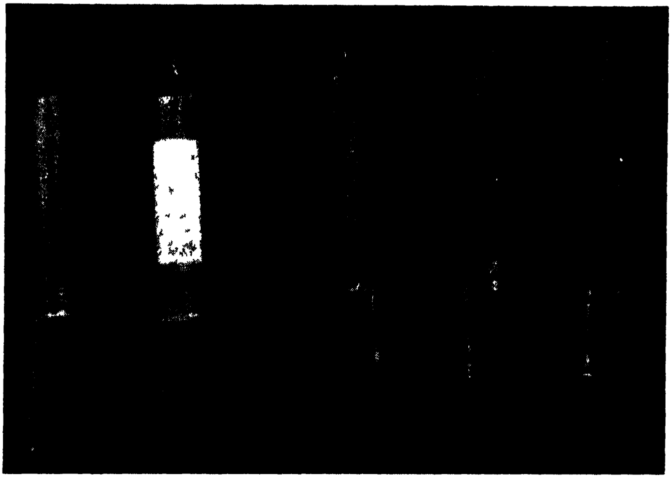


FIG. 6.—Reading left to right, uncoated Si-Ni alloy cathode sleeve, carbonate coated sleeve, Ba_2SiO_4 interfaces on Si-Ni sleeves.

The fact that the interface layer is easily seen and produces good x-ray diffraction patterns as well as details of its electrical behavior, led[51] to the prediction of a probable thickness range of 10^{-4} to 10^{-5} cm. An x-ray technique[58] is now available for measuring film thickness in this range provided the absorption coefficient and the crystal structure factor of the material is known. The diffraction pattern of thin interface films consists of diffraction lines from the film itself as well as lines from the underlying nickel. A comparison of the relative intensities of diffraction lines from the two materials may be used to calculate the thickness of the interface layer.

This method has been applied only to the silicate interface[33] yielding the results shown in Table II. The thickness at zero hours was measured immediately following the conversion process. Cathodes were heated in a quiescent state, without drawing emission current, to form the interfaces examined at 50 and 100 hours. The increase in interface thickness following conversion probably results from a silicon-BaO reaction. Ionic transport of barium to the base metal is generally considered to occur when emission current is taken from the cathode. The effect which this process has on the rate of growth of the interface thickness has not been investigated. Thickness measurements made on d.-c. operated silicate interface cathodes at the "end of life," 500 to 1000 hours, showed only a slight increase in thickness over that present at 100 hours. It is doubtful that the interface thickness alone is of prime consideration in bringing about the "end of cathode life."

TABLE II. Thickness of Ba_2SiO_4 interface formed on 4% Si-Ni alloy by (BaSr)O coating at 1150°K.

Time (hrs.)	0	50	100
Thickness (cm.)................	4×10^{-4}	1.4×10^{-3}	2.2×10^{-3}

The mechanism of interface formation involves the diffusion of silicon, of BaO, or both, hence the coating-interface boundary need not be sharply defined but rather is a region of variable composition of the silicate and the oxide. All interfaces which have been examined are firmly bonded to the base metal in contrast to the loose bonding which usually exists between the coating and interface.

3. *Conductivity*

Passage of an electron current density j through the interface produces a voltage drop across this layer of jt/σ, σ being the effective specific conductivity and t the thickness of the interface layer. Measurements of this interface voltage by means of probes imbedded in the coating will be discussed in a later section. These voltages can be used to evaluate σ only if the interface thickness is known or is estimated.

The effective specific conductivity of Ba_2SiO_4 and its temperature dependence were measured[33] using the technique employed in the oxide conductivity studies. A sample of the silicate, synthesized from $BaCO_3$ and SiO_2, was coated onto a ceramic rod. After an outgassing in vacuum following a heat treatment in carbon dioxide to remove the carbon residue from the binder, a series of conductivity measurements were taken, Fig. 7. Curve 4 represents the initial state following the removal of the carbon dioxide and the slope of the high temperature section indicates a $\Delta\epsilon$ of $5.2eV$. Continued heat treatment in vacuum increases the conductivity without a marked change in activation energy, curve 5, and passing a high current density through the sample further increased the conductivity, curve 6. A slight conductivity decrease followed re-exposure to carbon dioxide, curve 7.

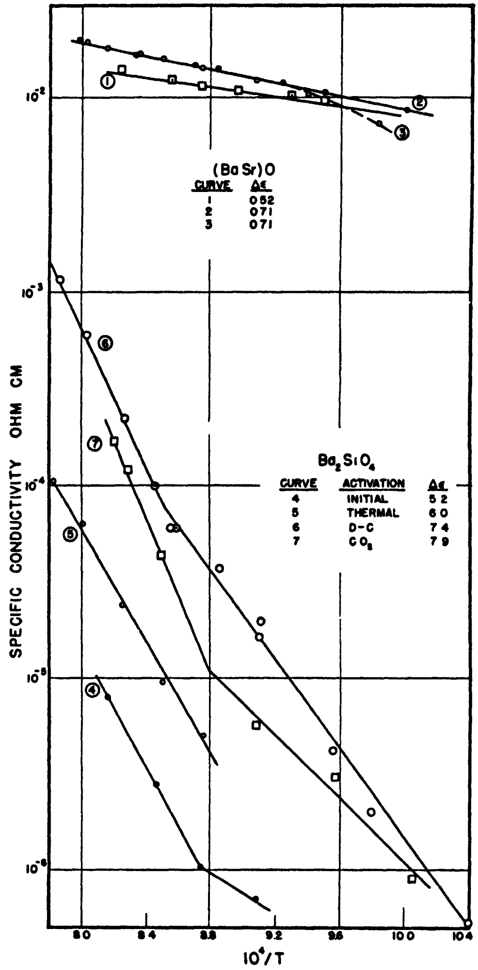


FIG. 7.—Effective electrical conductivity of (BaSr)O and Ba_2SiO_4 samples showing variation with reciprocal temperature and state of activation.

In general, the conductivity behavior of the silicate resembles that of the oxide regarding its reaction to activating and deactivating treatment. The slight "break" in the curves suggests two sources of electrons and two activation energies, although the data does not warrant extensive deductions of this sort. The most striking feature of this set of curves is the apparent high activation energy, 5 to 8*eV*. Were it not for the conductivity dependence on activation, the high activation energy would suggest the behavior of an intrinsic semiconductor.

Curves 1, 2 and 3 of Fig. 5(A) are reproduced in Fig. 7 for comparison with the interface conductivities. Using curve 6, as typical of the activated interface conductivity, and interface and coating thickness of 10^{-3} and 10^{-2} cm. respectively, certain deductions can be made from this figure. At no temperature less than about 1300°K. will the voltage drop across the coating equal that appearing across the interface and at 1000°K. the interface voltage-coating voltage ratio is 100:1. A current density of 10 $A/\text{cm.}^2$ passing through a cathode at 1150°K. should give rise to a 200 volt drop across the interface and a 10 volt drop over the coating. These values, as we shall later see, are consistent with the potentials developed within the cathode under pulsed conditions. Reducing the operating temperature should rapidly increase the interface voltage, this also being consistent with experimental measurements.

IV. PROPERTIES OF THE COMPLETE CATHODE

In this section we shall consider those properties of the cathode which are only studied using the complete cathode, base metal, interface and oxide, and may be subject to the influence of more than one of these components.

1. *Thermionic Emission*

Although the oxide coated cathode exhibits many interesting physical characteristics, the considerable expenditure of effort already devoted to studies of these properties is due to but one, the cathode's ability to supply a copious thermionic emission of electrons at a relatively low temperature. Neglecting the interface for the moment, Fig. 8(A) represents a cross section view of the oxide in contact with the base metal and Fig. 8(B), the energy level configuration of this model. Fig. 1 has been modified, only by joining the chemical potential of the semiconductor, μ, with the Fermi level of the metal, W_i. The Fermi level represents the highest occupied energy state of the metal at absolute zero and enters into the Fermi distribution function for the metal in a manner analogous to the chemical potential for semiconductors, eq. (2). Equating the rate of transport of electrons of energy ϵ from the conduction band of the semiconductor into the metal and the rate of transport of electrons of the same energy from the metal to the semiconductor leads to the result,

$$W_i = \mu \tag{18}$$

Schottky and Rothe[59] and Schottky[60] have considered the thermionic current density which may be obtained from the cathode described by

Fig. 8. A Richardson type of equation is found,

$$j_0 = A(1 - r)T^2 e^{-\frac{e\phi}{kT}} \tag{19}$$

where j_0 is the "saturation" current density,

$$A = \frac{4\pi m^* k^2 e}{h^3} = 120\ A/\text{cm.}^2\ \text{deg.}^2 \tag{20}$$

r is the mean reflection coefficient for electrons at the emitting surface, and ϕ is the total thermionic work function, $(\chi + \epsilon_0 - \mu)$. Substituting the value in eq. (8) for $\epsilon_0 - \mu$ we have the more familiar[1] thermionic emission expression,

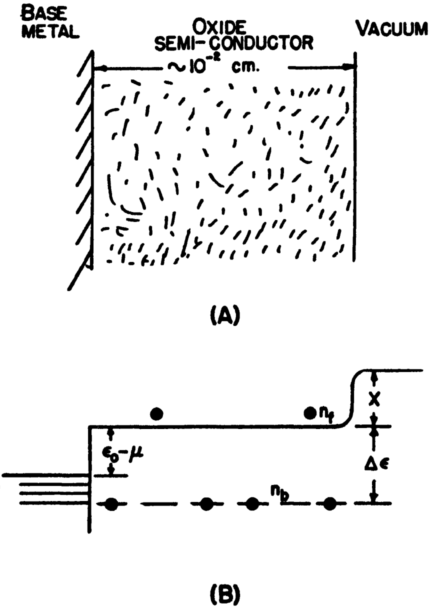


FIG. 8.—(A) Cross section of interfaceless cathode; (B) energy levels for this model.

$$j_s = B(1 - r)n_b^{\frac{1}{2}}T^{\frac{5}{4}}e^{-\frac{\chi + \Delta\epsilon/2}{kT}} \tag{21}$$

and

$$B = \frac{\sqrt{2}\, e(2\pi m^* k)^{\frac{1}{4}}}{h^{\frac{3}{2}}} \tag{22}$$

A Richardson plot of $\ln j_0/T^2$ vs. $1/T$ should not yield a straight line due to the temperature dependence of the total work function ϕ. Actually, no experimental measurements of the saturation current have been made with sufficient accuracy nor over a large enough temperature range to show this nonlinearity in the Richardson plot.

Measurements of the work function of oxide coated cathodes are frequently subject to criticism because of the particular method chosen to define the "saturation" current. The point of departure from a space charge limited condition is much less clearly defined than for pure metals. Nevertheless, most recent determinations of the work function of a (BaSr)O coated cathodes result in values near 1*eV*. Nishibori and Kawamura[61] report values of the thermionic work functions of SrO and (BaSr)O as given in Table III (see section IV-5).

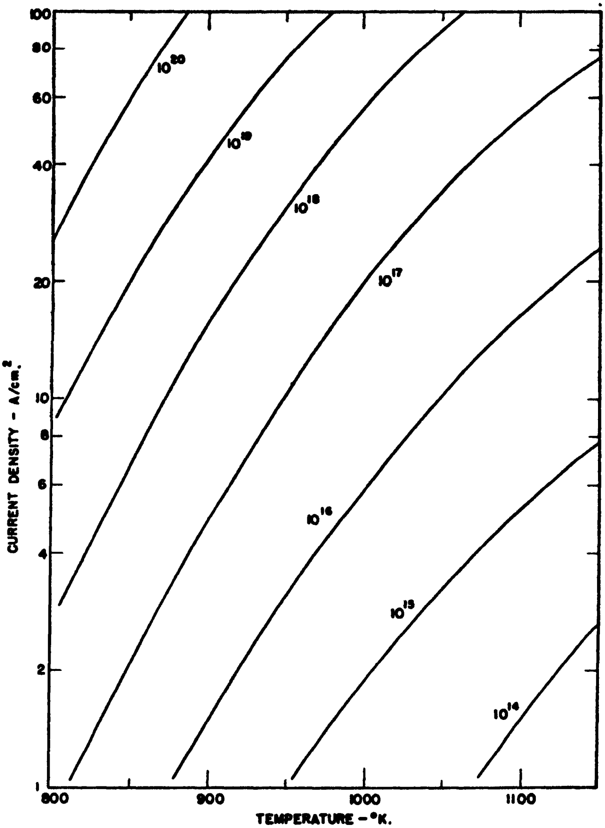


Fig. 9.—Temperature dependence of emission, eq. (21) for several values of n_b computed for $\phi = 1eV$, $r = 0$.

The presence of two unknown quantities, r and n_b, in eq. (21) prevents the determination of either from a Richardson plot alone. Nevertheless, speculations concerning possible values of r are not entirely unwarranted. For a uniform surface, the fraction of electrons with normal incident, energy components greater than $-e\Phi_v$, Fig. 1, which do not escape, should be small and thus r will be small. The presence of a potential hump at the surface instead of the step barrier would give r a value near unity since the probability of electron transmission through such a barrier is very small. Making use of the emission data of Prescott and Morrison[41] and their value[1] of 10^{20}/cm.^{3} for n_b and following the type of calculation made by Blewett,[1] eq. (21) gives the transmission coefficient, $(1 - r)$, the value 5×10^{-2} and $r = 0.95$. If instead, Wooten's[48] upper limit value of 10^{18}/cm.^{3} is used for n_b, the transmission coefficient becomes 0.5 and $r = 0.5$. Thus there is some justification in suggesting that the transmission coefficient is not necessarily small, as it has been considered in the past.

If the transmission coefficient is unity and the total work function is $1.0eV$, Fig. 9 represents the temperature dependence of emission for several values of n_b, as given by eq. (21). Allowing n_b a value between 10^{17} and 10^{18}/cm.^{3} leads to emission densities considerably in excess of the usually accepted d.-c. emission capabilities of an oxide cathode. Coomes[27] reports microsecond pulsed emissions of between 40 and 100 A/cm.^{2} at 1100°K. from oxide cathodes. Similar cathodes were found to have a work function of about $1eV$, as determined by a Richardson plot of pulsed emission measurements. Sproull's[62] pulsed cathodes showed an initial emission whose temperature dependence indicated a work function of $1.12eV$. Thus it seems not too unreasonable that the short time, pulsed emissions which may be taken from oxide cathodes and were heretofore called "enhanced pulsed emissions," represent the true emission characteristics predicted by eq. (21). The usual d.-c. emissions would then represent a lower emission state brought about as some consequence of drawing emission current. This possibility will be considered again in section IV-8. Further studies of the reflection coefficient, emission decay and the proper evaluation of n_b are necessary before the emission equation can be adequately tested.

The presence of an interface layer considerably modifies the cathode model of Fig. 8. Fig. 10(A) represents a cross section view of the interface type cathode whose energy level configuration is shown in Fig. 10, (B) and (C). (B) is the equilibrium state for zero current flow and in (C), the applied anode voltage V_A allows the flow of an emission current through the interface and coating, producing a voltage drop V_{ic}. In this figure the barrier layer is considered to be an N-type, impurity semiconductor, although an intrinsic semiconductor could be similarly represented. The chemical potentials of the interface and the oxide are equated to the Fermi level in the metal by balancing the rate of electron transport through the cathode. The height of the barrier will then be approximately $\Delta\epsilon'/2$ with $\Delta\epsilon'$ the thermal activation energy of the interface material. Thermally excited electrons from the metal pass over the barrier, are transported through the interface and coating, and emitted from the external surface. The interface barrier may exceed the surface barrier in height, as shown in this figure. Electrons entering the oxide from the interface with an appreciable kinetic energy will dissipate this energy after a few interactions with the lattice and will proceed in the normal conduction band. Drawing emission current causes a tipping of the energy levels to an extent determined by the electrical conductivity of the material.

The current density of electrons emitted from the external surface will be governed by the electric field at this surface and will depend on

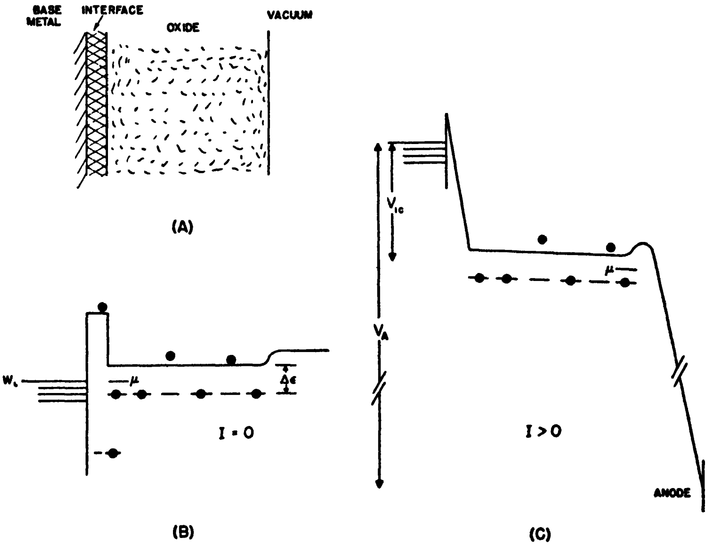


FIG. 10.—(A) Cross section of interface type cathode; (B) energy levels with no emission; (C) energy levels when emission current is drawn.

the interface only insofar as the interface reduces this field. When the current is limited at the interface barrier or by the interface conductivity, an appreciable fraction of the applied anode-cathode voltage will appear in this region and thus reduce the electric field at the external surface. Thus, for a given anode voltage, the presence of an interface barrier layer may reduce the emission current but should not influence the "saturation" current j_0, directly. Indirectly, chemical reactions at the base metal could conceivably influence the excess barium content of the cathode, and thus influence n_b.

The voltage-current characteristics of a barrier layer have been considered theoretically, and good agreement is found between experiment and theory, particularly for blocking layer rectifiers of the copper oxide type. Wright[51] has extended Mott's theory[63] to oxide cathode interfaces with considerable success, explaining the electrical characteristics of probes imbedded in the coating. This theory is applicable if the interface is considered to be an insulator or to have a much lower conductivity than the oxide. In the absence of a "chemical" barrier, i.e., a layer whose composition differs at least stoichiometrically from that of the coating, a Schottky barrier may exist. In a series of papers published between 1939 and 1942 Schottky developed a barrier layer theory to explain the action of the selenium rectifier which may not contain a "chemical" barrier. A potential barrier is formed at the metal-semiconductor boundary due to a space charge formed in the semiconductor when equilibrium is established between the two materials. The theories of both Mott and Schottky place the barrier height equal to the difference between the work functions of the metal and the semiconductor. In Schottky's theory, the thickness of the space charge barrier will vary with the current density whereas the interface has a fixed thickness in Mott's derivation. An excellent review of Schottky's papers prepared by Joffe[64] may be consulted for the details of this mechanism. Mott's theory is discussed in section IV-6.

2. Photoelectric Emission

Modern semiconductor theory predicts that the photoelectric work function should exceed the thermionic work function. Measurements of the photoelectric emission from oxide cathodes, made at near room temperature, should exhibit a long wavelength limit at a photon energy equal to $\chi + \Delta\epsilon$. At this temperature relatively few electrons will be thermionically excited to the conduction band and the photon energy will be expended in exciting electrons located at impurity centers. In order that they may be emitted from the oxide, at a temperature at which thermionic emission is negligible, the electrons must be raised to the top of the conduction band by acquiring an energy $\chi + \Delta\epsilon$ (see Fig. 1). For comparison, the thermionic work function of the cathode is $\chi + \Delta\epsilon/2$. eq. (21).

Huxford's[65] measurements of the photoelectric and thermionic work functions of the same cathode yield nearly similar values, in disagreement with the above theory unless it is assumed that $\chi \gg \Delta\epsilon$. This discrepancy is removed in a more recent determination of the photoelectric threshold made by Nishibori, Kawamura, and Hirano.[66] Fig. 11 shows the photoelectric emission (corrected for variations in the incident intensity which depend on wavelength) from three cathodes as a function of the energy of the incident photon. Linear sections of the curves were extrapolated to the axis to determine the photoelectric threshold values given in Table III. Values shown for the thermionic work function, the thermal activation energy and the electron affinity were obtained from a separate experiment. Although good agreement is obtained between the photoelectric work function and the other equivalent values, these results would be more satisfying had they all been obtained from the same cathode. The extent of the "foot" of the curves of Fig. 11 was found to be temperature dependent and was explained as a thermionic emission of electrons which were excited into the conduction band by photons of insufficient energy to carry them to the top of the band. The previous extrapolation simulates the photoelectric emission at a temperature of absolute zero.

The similar values of photoelectric work function for BaO and (BaSr)O were taken[66] to indicate that the difference in thermionic emission capabilities of the two materials, Fig. 3, results from a higher equilibrium value of n_b in the solid solution than in the pure oxide. This must be regarded as speculative until a value for χ or $\Delta\epsilon$ alone, has been obtained for BaO.

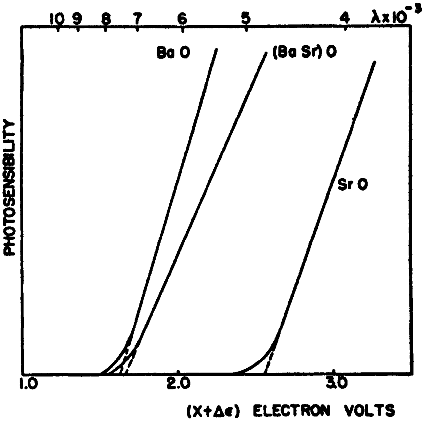


FIG. 11.—Photoelectric current from oxide cathodes as a function of incident photon energy.

3. *Emission in Retarding Fields*

The distribution in energy of electrons emitted from a cathode may be studied using a magnetic or electrostatic velocity selector, or more easily by measuring the diminution in anode current as the anode is made increasingly negative. In this retarding potential region only those electrons will reach the anode which have energies greater than the height of the barrier imposed between the cathode and anode by making the anode negative. Contact potentials within the diode as well as the applied retarding potential are effective in establishing this barrier height. For plane-parallel electrodes, a current density j_R is collected at an applied anode-cathode voltage V and,

$$j_R = A(1 - r)T^2Ce^{-\frac{e(V+\phi_A)}{kT}} \tag{23}$$

where ϕ_A is the anode work function, C is a factor representing electrons which are reflected at the anode surface and re-enter the cathode, and

A and r have values described in eq. (19). When cylindrical geometry is used, the equation assumes a similar but more complex form, and if the anode diameter to cathode diameter ratio is large the anode reflected electrons can be neglected. Eq. (23) is valid only for applied voltages which cause a potential barrier greater in height than the normal work function barrier, ϕ, in Fig. 1. For somewhat more positive values of V, the collected current has a constant value j_0 given by eq. (19). A plot of ln (j_R/j_0) vs. V should yield a line of slope $-e/kT$ which abruptly joins a horizontal line at $V + \phi_A$ equal to ϕ.

Many of the earlier attempts to check the temperature of an oxide cathode from the slope of the retarding potential curve were without success.[1] More recently Heinze and Hass[67] and Fan[68] were able to obtain agreement with the predicted slope.

In case the anode work function is known, the applied voltage at which the "break" in the curve occurs may be used to evaluate the work function of the cathode. Champieux[69] used a tungsten filament for the anode and by flashing this to a high temperature he felt justified in assuming the work function to be that of clean tungsten. The geometry of such a diode would undoubtedly be very poor. Heinze and Wagener[70] determined ϕ_A for their anode by measuring the emission from an auxiliary tungsten filament in retarding fields. The work function of the tungsten filament was found from the slope of its Richardson plot. True work function values can be measured by this method *only* if a simple step barrier, such as shown in Fig. 8, exists at the surface. The presence of a potential hump will invalidate any measurements of this quantity involving the emission of electrons. Only the vibrating electrode technique,[71] the Kelvin method, can be relied upon for measuring the true work function of the oxide cathode surface. Patch effects, arising from exposed crystal faces of differing work function, must be carefully considered in the interpretation of these phenomena.

Although the argument cited above places in question values of the work function obtained using retarding potentials, a comparison of the data with eq. (23) should at least permit an evaluation of $(1 - r)$, the transmission coefficient. Heinze and Wagener[70] show that over a considerable range of activation, cathode emission obeys eq. (23) and the transmission coefficient has a value near unity.

The shape of the retarding potential curve in the vicinity of the "break" has been the subject of much controversy. Low cathode temperatures are used so that space charge effects do not modify the curve. If the curve shows a sharp "break" and the slope has a value corresponding to the true cathode temperature, the conclusion is reached that the energy distribution is Maxwellian. A retarding potential plot by Fan,[68] Fig. 12, was interpreted in this manner. The presence of a large reflection coefficient at the oxide surface, due perhaps to patch fields or to a thin penetrable potential barrier, would be expected to reduce the proportion of low energy electrons reaching the anode. This behavior is frequently observed in measurements on oxide cathodes and is shown in Fig. 13. In these results obtained by Brown[72] the absissa is V/T so that curves taken at various temperatures should coincide if reflections are missing. Absence of such coincidence was interpreted as implying a very high reflection coefficient.

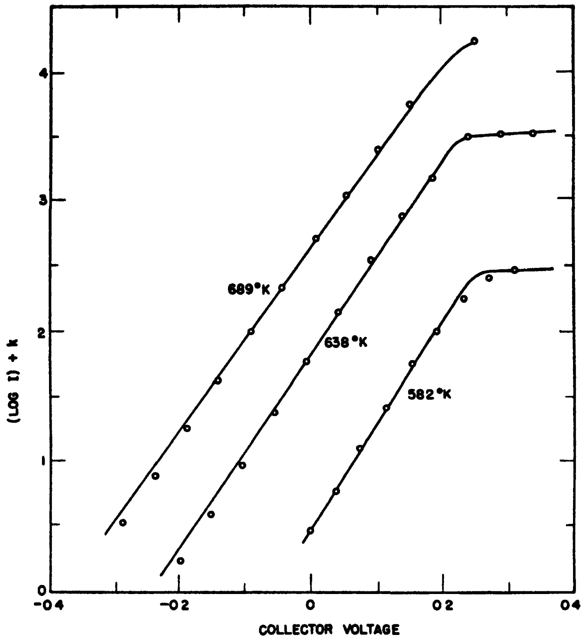


FIG. 12.—Electron emission from a BaO cathode in a retarding field. Slope of solid line through points computed from eq. (23) using the measured temperature of the cathode.

Thus, considerable discrepancy exists among the results of various studies of emission in the retarding potential region. The removal of these points of dissention should provide a means for examining the nature of the surface barrier as well as evaluating the true work function.

4. Emission in Accelerating Fields

Positive potentials applied to the diode anode will provide an increasing anode current, first in accord with the Langmuir-Child space charge relationship and at higher voltages, following a Schottky type of behavior. Emission in the Schottky region is generally compared with the characteristic predicted for a clean metal surface. The presence of an electric field intensity E at the metal surface, effectively reduces the work function ϕ by an amount,

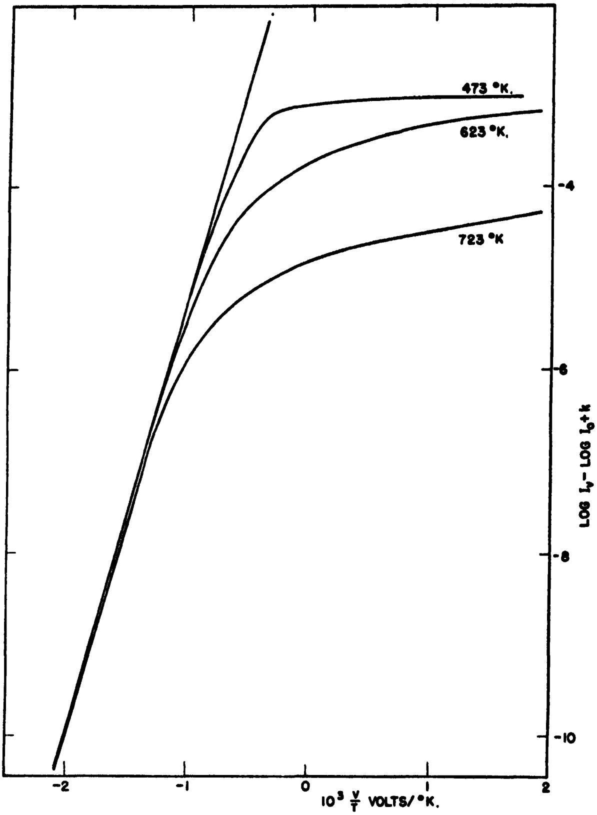


FIG. 13.—Electron emission from (BaSr)O cathode in a retarding field showing the deficiency of low energy electrons.

$$\Delta\phi = \sqrt{eE} \tag{24}$$

A plot of ln j_s vs. $E^{\frac{1}{2}}$ should yield a straight line for emission in the Schottky region where j_s is the measured anode current.

An example[69] of this is seen in Fig. 14, and typical of other measurements in this region, shows a linear relationship but a slope several times that predicted by the Schottky theory for metal surfaces. Similar slopes are found in both d.-c. and pulsed[27,62] emission measurements. A semiquantitative explanation of this incorrect slope was presented by

Rose[73] who made use of the patch effect. More recently, Morgulis[74] derived a Schottky effect equation for a semiconductor emitter assuming the field to distort the energy levels near the surface and found a reasonable agreement with the Schottky plots obtained by Sproull.[62] When a fit is made with the experimental curves a value for n_f of about $10^{15}/\text{cm.}^3$ was obtained from the theoretical equations. An extrapolation of the Schottky plots to zero field, see Fig. 14, permits an evaluation of the zero field emission from the cathode. In the absence of a potential hump at the surface, this method for determining j_0 is preferred since the emission from an oxide cathode seldom shows a clear departure from space charge limited conditions.

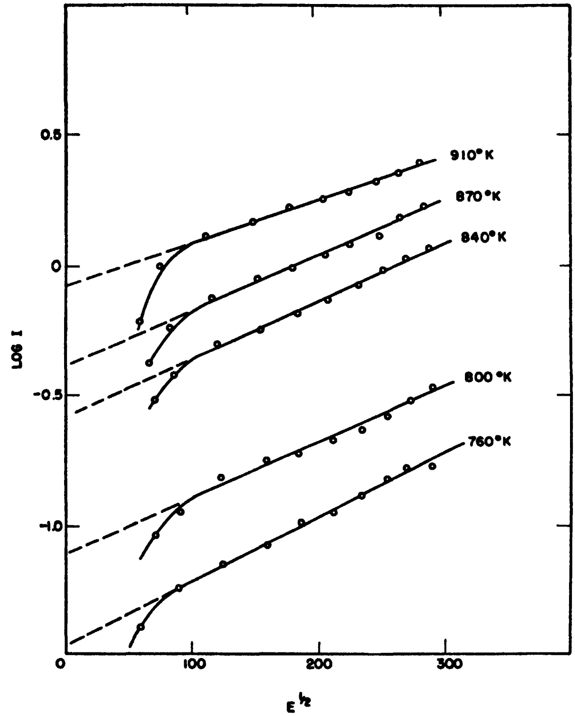


Fig. 14.—Variation of electron emission in the Schottky region as a function of the square root of the electric field intensity.

A wave mechanical treatment of the flow of electrons over a barrier predicts that interference phenomena should occur. The transmission coefficient will be a function of the wavelength of the electron, and thus its energy, as well as the shape of the barrier. Although the effect is small and easily overlooked, its occurrence has been established for the emission of electrons over the work function barrier of clean metal surfaces.[75,76] Maxima and minima are found in the Schottky plot whose position and magnitude agree with the theoretical predictions. Champieux[69] has presented the data shown in Fig. 15 as evidence for periodic deviations in the emission of electrons from oxide coated cathodes. The deviations are very small indeed and unless the curves were completely reproducible an objection might be raised on the ground that instabilities in oxide cathode emission are of about the same order of magnitude.

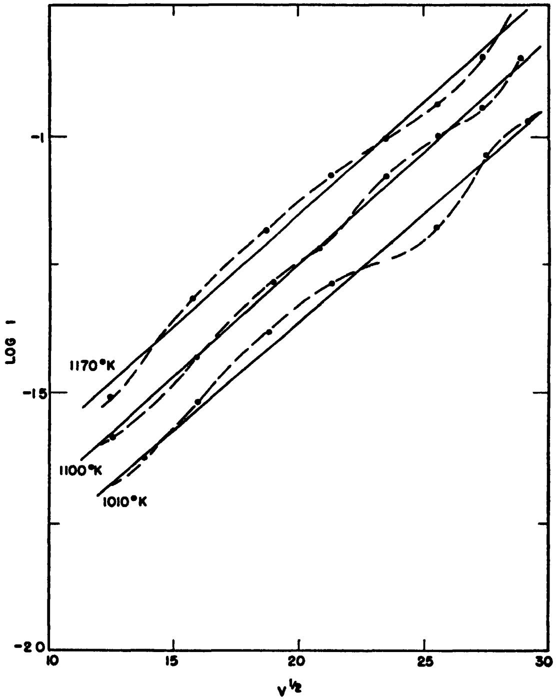


FIG. 15.—Electron emission in an accelerating field showing possible periodic deviations.

Should the presence of periodic deviations be confirmed their location and shape may contribute information regarding the barrier shape, be it at the surface or at the interface.

5. *Conductivity*

Nishibori and Kawamura[61] made use of a probe wire imbedded in the cathode coating to study the temperature variation of conductivity. When an emission current is drawn from the cathode the probe assumes a positive potential with respect to the base metal. This potential is determined by the geometry, the emission current and the conductivity of the material lying between the probe and base metal. Since both the interface and a part of the coating are located in this region some care is required in interpreting such conductivity results.

D.-c. measurements of the conductivity of cathodes prepared on a pure nickel base metal probably contain only a small contribution from the interface region. Measurements of this type were made by Nishibori and Kawamura. The temperature dependence of conductivity was interpreted by means of eq. (7) to evaluate $\Delta\epsilon$, values of which are found in Table III. Combining these values of $\Delta\epsilon$ with the thermionic work functions previously discussed, allows an evaluation of χ. This surface work function is seen to be small and from additional values contained in the original paper, the surface barrier seems to be independent of the state of cathode activity. As χ has nearly the same value for SrO and (BaSr)O, this may be interpreted as representing the same surface conditions on both coatings, i.e., pure SrO as seen in section II-3.

TABLE III. Photoelectric work function, $\chi + \Delta\epsilon$; thermionic work function, $\chi + \Delta\epsilon/2$; thermal activation energy, $\Delta\epsilon$; and electron affinity, χ; for oxide cathode materials obtained by Nishibori, Kawamura, and Hirano.

	$\chi + \Delta\epsilon$	$\chi + \Delta\epsilon/2$	$\Delta\epsilon$	χ
BaO	1.63			
(BaSr)O	1.66	.98	1.4	.28
SrO	2.58	1.37	2.1	.32

Measurements of both conductivity and thermionic emission during the activation of a cathode has resulted in an interesting relation between the two. Fig. 16 shows a straight line drawn through three points reported by Nishibori and Kawamura. The interpretation of this relationship may be sought by comparing eq. (10) showing the effect of activation on the conductivity and eq. (19) relating the saturation emission current density to the total work function. Equating these expressions through the common term exp $[(\mu - \epsilon_0)/kT]$ results in the ratio,

$$\frac{j_0}{\sigma} = \frac{3(1 - r)kT}{4l_0e} e^{-\frac{\chi}{kT}} \tag{25}$$

Only if χ is independent of a change in cathode activity, should eq. (25) result in a linear relationship between j_0 and σ. The electron affinity χ is directly related to the surface dipole moment which in the earlier emission theories governed the emission process. The invariance of χ over three orders of magnitude of emission current offers a definite point of conflict with any theory which defines the emission process as only a surface phenomena. Similar experimental results were recently reported by Hannay[77] although the details of these experiments are not as yet published.

The double probe technique has served as a valuable tool in separating the interface and coating conductivities. Two probes imbedded in the coating at different depths permit measurements of the potentials appearing at these positions when an emission current is drawn. An extrapolation of this potential to the interface depth results in an excess voltage which is attributed to the interface conductivity. The difference of potential between probes is used to evaluate the coating conductivity. This technique with microsecond pulses was used to examine the potentials which are developed across the coating and interface of the cathode over a wide range of emission current densities.[52]

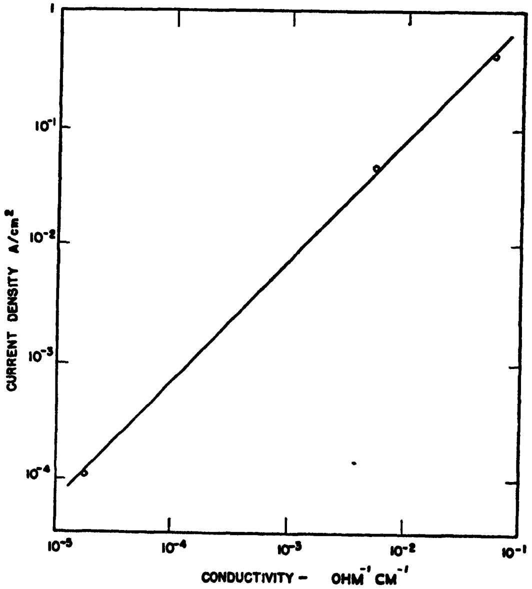


FIG. 16.—Correlation of electron emission current density and the coating conductivity, see text.

In a cathode prepared on a pure nickel base, the interface, which probably contributes a negligible amount to the d.-c. measured conductivity, may contribute appreciably to the conductivity measured with pulses of high peak current density. The interface thickness is usually unknown in experiments of this type, hence voltages developed across the interface and across the total coating are compared rather than the respective specific conductivities. Changes in the interface voltage to coating voltage ratio which occur with changing current density are due to the nonohmic nature of the interface contact as compared to the ohmic conductivity of the coating. Evidence for the nonohmic interface contact is presented in section IV-6. Thus, at low current densities the interface voltage may be negligible whereas at high current densities it may be comparable to the coating voltage.

Fineman and Eisenstein[52] described the results of double probe pulsed measurements on one pure electrolytic base cathode which at 10 $A/\text{cm.}^2$ had an interface voltage in excess of the coating voltage. More recent measurements by Dillinger[78] and Mutter[79] place the interface voltage at about one-half the coating voltage when measured under similar conditions and explain the above cited high ratio as due to a poor mechanical coating bond.

Both d.-c. and pulsed methods of measurement have been used to detect interface voltages in cathodes which have a "chemical" interface barrier layer. Wright[51] prepared cathodes on a magnesium-nickel alloy base metal and used an aluminum-nickel probe wire in the coating for d.-c. measurements of conductivity at relatively low temperatures. The voltage-current characteristics of the probe were interpreted by means of Mott's theory to indicate the presence of barrier layers on both the probe and base metal surfaces. Presumedly these were due to the aluminate and MgO interfaces mentioned earlier. The effective conductivity appearing between the probe and base metal seemed independent of the probe-base metal spacing and therefore was attributed to the interfaces rather than the coating. The temperature dependence of this conductivity was used to evaluate the thermal activation energy, $1.2eV$. Further details of the voltage-current plot were believed to show a value of $0.7eV$ for the height of the interface barrier.

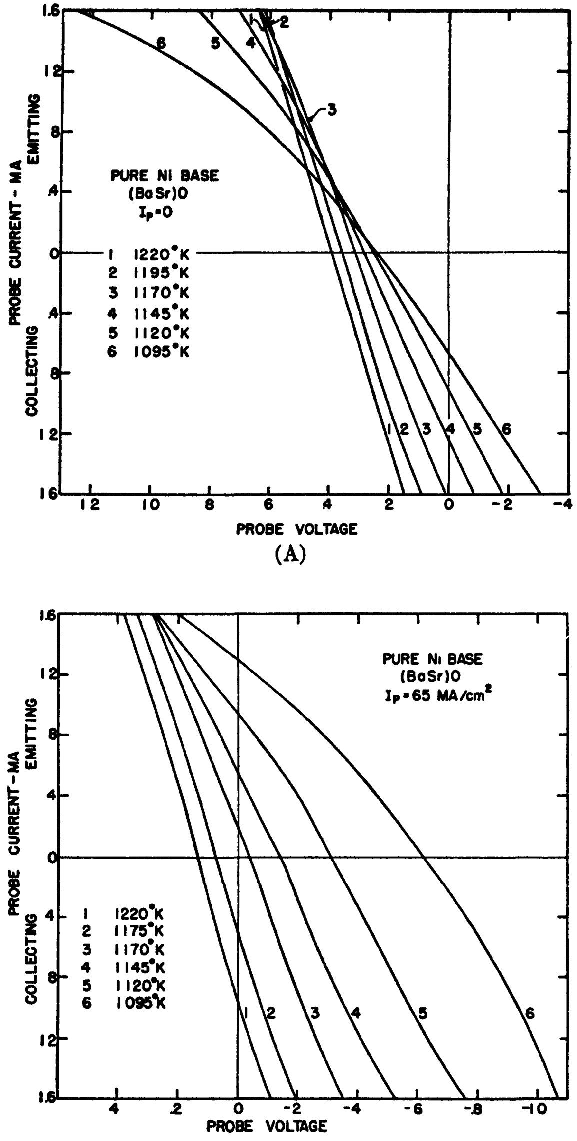
Measurements[51] in which the probe wire was used as a null current, potential indicator showed a saturation of the coating conduction current at about the same current density as that at which saturation of the thermionic emission took place. Approximately one-tenth the applied anode voltage appeared between the base metal and probe in these experiments. If this potential drop occurs primarily at the interface, this observation would seem to indicate that the emission limitation exists at the interface rather than at the external surface. In future studies of this nature, the use of double probes should remove any question regarding the exact location of this emission limitation.

Fineman and Eisenstein[52] examined the voltages appearing in the coating of cathodes which had a chromium interface and concluded that at 1125°K. the interface voltage may be five times the coating voltage for an emission of 10 $A/\text{cm.}^2$ but less than the coating voltage at 0.5

$A/\text{cm.}^2$. This again is due to the nonohmic character of the interface contact. Increasing the cathode temperature decreased the interface voltage more rapidly than the coating voltage so that at 1225°K. the voltages at 10 $A/\text{cm.}^2$ were nearly the same. A possible explanation for this is found in the thermal activation energies of the two materials which cause the interface conductivity to change more rapidly with temperature than does the oxide conductivity (see Fig. 7). Results similar to those discussed above are obtained in cathodes containing a Ba_2SiO_4 interface and are described in section IV-6.

Null current, probe measurements of electrostatic potential are subject to criticism unless the probe characteristics are known. Probes may disturb the true conditions of the coating, they may show polarization effects and in the presence of both electronic and ionic conduction the null current potential must be interpreted with care. Polarization effects are absent if platinum probes are used and the thermionic emission properties, characteristic of the type of base metal used, seem unchanged by the introduction of probes. Mutter[79] has investigated the d.-c. characteristics of a platinum probe imbedded in a (BaSr)O coating on an electrolytic nickel base metal. See Fig. 17. For zero emission current, the characteristics in *A* show no change of slope on crossing the zero current axis. The influence of an interface contact, probably at the probe, is seen in the asymmetry of the curves about the zero current axis, particularly at the lower temperatures. An emission current density of 65 ma/cm.^{2} caused a displacement of the curves by an amount interpreted to be the change in oxide potential resulting from the current flow. The following two methods of conductivity measurement were compared. The slope of the curves, $\Delta I/\Delta V$ at zero probe current, defines one conductivity. When an emission current is drawn, the probe voltage displacement from its value at zero emission current gives the potential drop produced in the cathode by the flow of emission current. The ratio of the emission current to this potential drop defines a second conductivity and the two values seem to be in rough agreement. The slopes of the curves at the zero current axis are not appreciably changed for the two conditions of emission current. A study of the probe characteristics under pulsed conditions, made by Dillinger,[78] led to similar results. Whenever it is possible to compare the results of probe measured potentials with values obtained by other means, a reasonable agreement is found.

Loosjes and Vink,[80] who criticized the probe technique, introduced a novel method for measuring the potential developed across the complete cathode when pulsed currents are drawn. A movable anode was employed to obtain the voltage-current plots shown in Fig. 18(A) for

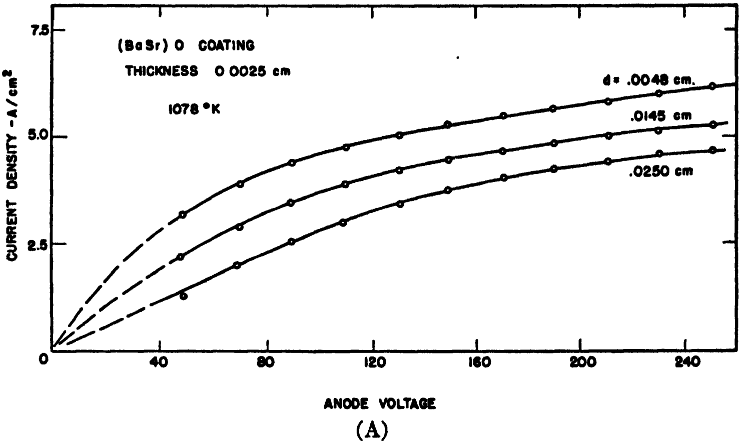


(A)

(B)

FIG. 17.—Probe voltage-current characteristics as a function of temperature, for two values of emission from the cathode.

three values of the anode-cathode surface spacing, *d*. Replotting this data, using the anode voltage and the spacing as coordinates for fixed values of emission current, allowed an extrapolation to the voltage which would exist at the cathode surface, i.e., at $d = 0$. The variation of this voltage with current is seen in Fig. 18(B). Curves obtained at several temperatures all exhibited the nonohmic behavior shown in this figure. A comparison of the plots (A) and (B) at a current density of 5 $A/\text{cm.}^2$ and at a spacing of 0.0048 cm., approximately twice the coating thickness,



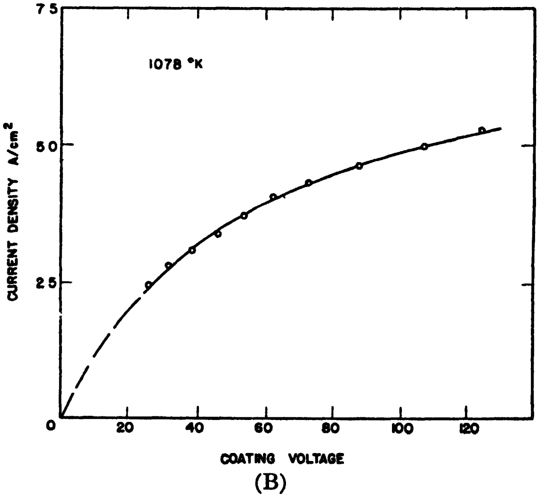


FIG. 18.—(A) Diode voltage-current characteristics for three values of the anode-cathode spacing. (B) Voltage-current characteristics of the coating determined from Fig. 18A, see text.

shows the striking result that of the 130 volts applied between anode and cathode 110 volts appears within the cathode. Certainly a large fraction of this is located in the interface region, as evidenced by the nonohmic characteristic.

The appearance of large voltages within the cathode, particularly across the interface layer, is probably associated with the sparking phenomena which will be described presently. Whether the observed interface voltage is due entirely to the low conductivity of the interface material or is due to a high field developed at the metal-interface contact to promote a more copius emission of electrons remains an open question. Values of the conductivity of pure Ba_2SiO_4 and values of a probable interface thickness give at least semiquantitative agreement with observed interface voltages.

6. *Rectification*

According to the theories of Mott[63] and Schottky[64] the presence of a barrier layer between a metal and a semiconductor should result in a rectification effect. If the semiconductor is of the N-type, the direction of low conductivity is from the metal into the semiconductor. An examination of Fig. 10 shows that the application of an electric field in the direction to induce electrons to flow from the metal into the semiconductor leaves the barrier height unchanged until the field becomes sufficient to promote Schottky emission or barrier penetration. Thus, no appreciable increase of current accompanies the increase of electric field. A field applied in the opposite direction raises the energy levels of the semiconductor relative to the levels in the metal and thus decreases the barrier height to current flow in this direction. An appreciable increase in current accompanies an increase in the electric field, resulting in a high conductivity. This explanation, based on Mott's theory, is modified by Schottky who allows the space charge layer to vary in thickness according to the direction of current flow.

Rectification at the interface barrier of oxide cathodes was demonstrated in the d.-c. voltage-current characteristics of an aluminum-nickel probe in a (BaSr)O coating, taken with respect to the magnesium-nickel base metal. Measurements made by Wright[51] on one such cathode are shown in Fig. 19. The direction of low conductivity is that in which electrons flow from the probe to the base metal. On the basis of the previous discussion this places the effective rectifying contact at the probe interface. This behavior would be expected even though barrier layers existed at both metal surfaces for the probe area is only $\frac{1}{65}$ the coated area of the base metal thus considerably increasing the current density at the probe interface.

Pulsed measurements, employing the double probe technique, were used by Mutter[81] in an ingeneous cathode structure to study rectification at the interface of two base metal types. Fig. 20 shows a cross section view of two flat cathodes facing each other, one prepared on a pure electrolytic nickel base and the other prepared on a 5% silicon-nickel

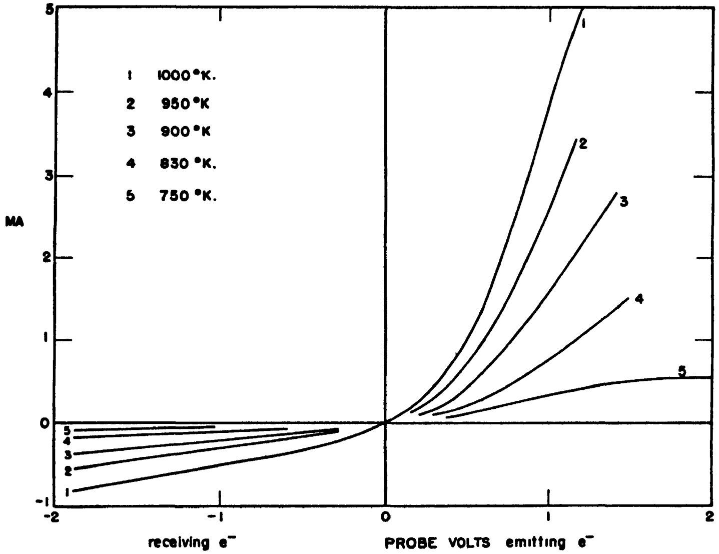


FIG. 19.—Probe voltage-current characteristics showing rectification at the interface on the probe.

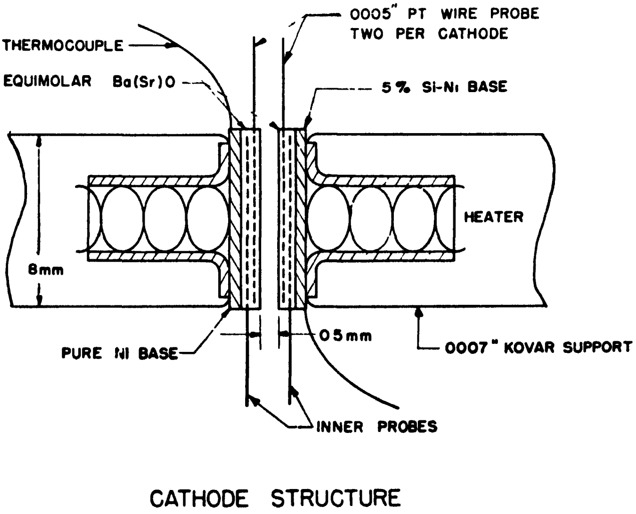


FIG. 20.—Double probe, double cathode structure used by Mutter, see text.

base. Two platinum probes imbedded in each coating allowed measurements of the interface and coating voltages with current passed in either direction through the cathodes. Results obtained at 1075°K. are shown in Fig. 21. The coating conductivity is seen to be ohmic and independent of the direction of current flow. Furthermore, the coating conductivity is independent of the presence of silicon in one base metal and not in the other. This is a particularly interesting result since for many years silicon and other reducing impurities have been added to the base metal, presumedly to increase the emission through the release of barium which according to our interpretation of eq. (6) should likewise increase the conductivity.

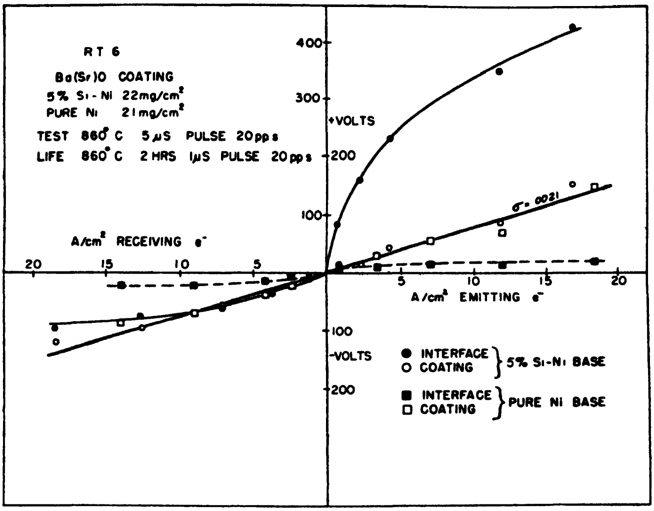


Fig. 21.—Voltage-current characteristics of the interface and coating on two types of base metal.

The interface voltage in the pure nickel base cathode is less than the coating drop and shows no change with the direction of current flow. In the case of the silicon-nickel base cathode, the interface voltage increases rapidly in the forward direction to a value four or five times the coating drop. In the reverse direction the increase is less rapid reaching a value nearly that of the coating drop. The direction of electron flow in normal cathode emission is the direction of low conductivity in accord with the theoretical predictions.

Changes which occur in the interface voltage with time are seen in Fig. 22. With increasing cathode life the pure nickel interface voltage increases and shows rectification properties, although the magnitude of this voltage remains less than the coating drop. Whether this interface voltage originates due to the build up of an interface compound or to a change in the oxide bounding the base metal is unknown. Certainly the pronounced rectification effects and the very high interface voltages are only found in the presence of a definite interface compound.

Throughout this discussion it was assumed that rectification occurred only at the metal-interface contact. It seems reasonable that some degree of rectification may also take place at the semiconductor-interface contact depending upon the type of interface layer which is present. Rectification at a metal-semiconductor boundary was discussed recently by Bardeen[82] in terms of surface states.

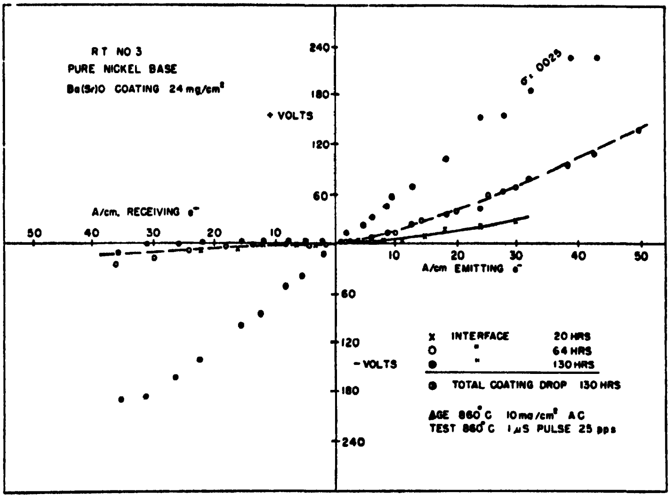


FIG. 22.—Voltage-current characteristics of the interface and coating, showing the effect of aging on the interface voltage.

7. *Thermoelectric Effect*

When two identical electrodes are imbedded in a semiconductor and a thermal gradient exists, producing a temperature difference ΔT between the electrodes, a potential difference known as the Seebeck emf will be established between the two junctions. This emf is essentially a function of the semiconductor since the Seebeck emf for metals is only about 0.1% as large as for semiconductors.

The Seebeck emf/degree at a temperature T has been related[83] to the fundamental properties of the semiconductor by the equation,

$$\frac{emf}{\Delta T} = -\frac{1}{e}\left(\frac{\Delta\epsilon}{T} + k \log \frac{n_f}{n_b}\right) \tag{26}$$

the sign being negative for an N-type, impurity semiconductor. A negative value for the complete expression implies that the electron current tends to flow from the metal to the semiconductor at the hot junction. The temperature dependence in this expression is determined by the relative magnitudes of the two terms in the bracket, eq. (26), the former causing the value to become less negative and the latter causing the value to become less positive with increasing temperature. Assuming $\Delta\epsilon$ to be about 1.0eV and using values of n_f and n_b shown in Table I, at 1000°K. the expression has a value of about -10^{-3} volts/degree. Due to the temperature gradient through the oxide coating, a probe near the surface should be negative with respect to the base metal.

The Seebeck emf was measured by Blewett[1] for a sample of BaO and found to be negative as well as of a magnitude not inconsistent with eq. (26). Nishibori and Kawamura[61] report a zero current, probe potential of -0.1 volts. A similar effect is seen in Fig. 17 indicating an increasing emf with higher temperatures which brings about a larger thermal gradient through the coating.

Future studies of the Seebeck emf as related to the other cathode parameters, $\Delta\epsilon$, n_b, and n_f, suggests a promising field for experimental research. The presence of an interface layer at the base metal will influence the emf only if a temperature gradient is present through the interface.

8. *Emission Decay Phenomena*

The emission current which may be taken from an oxide cathode is not always stable with respect to time, particularly when emission limited current is drawn. Fig. 23 shows the emission decay observed by Blewett[34] for a series of cathode temperatures. At still lower temperatures the decay may occur over a period of hours but eventually a stable emission level is reached. This is a repeating phenomena, the cathode recovering its initial emission when allowed to glow without drawing anode current. The rate of decay increases with increasing temperature and at normal operating temperatures the decay may not be detectable using d.-c. methods of measurement.

With the introduction of microsecond pulse methods of emission measurement, it was found that the saturation current which could be taken from a cathode in pulses considerably exceeded the usual d.-c. emission capacity. Sproull[62] attempted to bridge the gap between pulse and d.-c. measurements and showed that decay effects persisted at the higher temperatures, see Fig. 24. It is generally presumed that the decay shown in the two figures represents one and the same phenomena, although this is not necessarily so. Blewett found the effect to be independent of the applied anode voltage and therefore termed it a volume effect, but its explanation has been sought in terms of a change in the surface state. Either the electrolytic removal of barium, which was favored by Sproull, or the electrolytic deposition of oxygen at the surface have been discussed as plausible explanations. Wright[51] found similar decay phenomena in the conductivity between the base metal and a probe in the coating. Since the direction of current flow determined the onset of decay and since no external surfaces were involved, the effect was explained as an interface behavior. Electrolytic flow of barium ions to the interface might easily modify the shape of the barrier or change the barium concentration in the interface layer. Quantitative agreement is obtained between Sproull's theory and the observed decay only if 5 to 50% of the total conduction current is ionic. In order to explain the independence of decay with coating thickness, it is necessary to assume the ionic transport confined to the surface of the individual crystals of the coating. These somewhat questionable assumptions are not required in Wright's theory although this theory remains as yet qualitative.

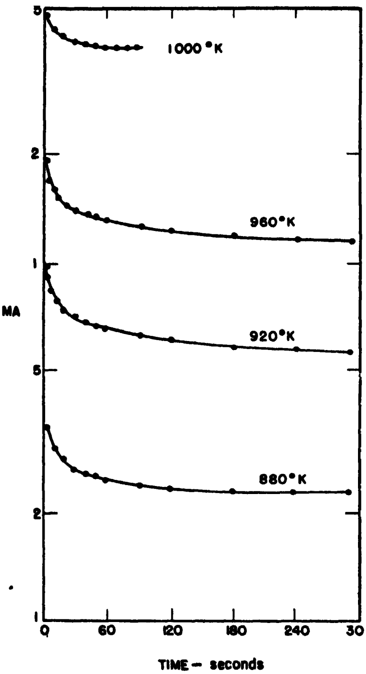


Fig. 23.—Long time emission decay at several cathode temperatures, (BaSr)O.

The results cited thus far seem to differentiate between d.-c. and pulsed emission in terms of a change in the physical state of the cathode, either at the external surface or at the interface barrier, resulting from an electrolytic phenomena. Certain additional factors should be considered.

1. Not all cathodes show the decay phenomena to the extent indicated by Figs. 23 and 24. Fan[68] observed no d.-c. emission decay in the time and temperature range shown in Fig. 25. Blewett[8] suggested later that different states of cathode activity may have accounted for this discrepancy. In an effort to correlate pulsed emissions in excess of 50 $A/\text{cm.}^2$ with the short time behavior predicted by Fig. 24, the writer has examined emission pulses of 10 μs duration for decay effects. At temperatures between 1000 and 1200°K., which was the only range studied, no perceptible emission decay was found. Ramsey[27] likewise found no microsecond decay in well activated cathodes.

2. The discrepancy between d.-c. and pulsed emission is not necessarily large. Microsecond pulsed emissions in excess of 100 $A/\text{cm.}^2$ at about 1100°K. are reported[27,33] from standard cathodes whose d.-c. saturation emission has been considered to be less than 1 $A/\text{cm.}^2$ Fineman,[85] recognizing that limitations on d.-c. emission were frequently imposed by anode effects, constructed a suitable test diode and was able to obtain a d.-c. emission at 1175°K. of 14 $A/\text{cm.}^2$ for several hours. The emission

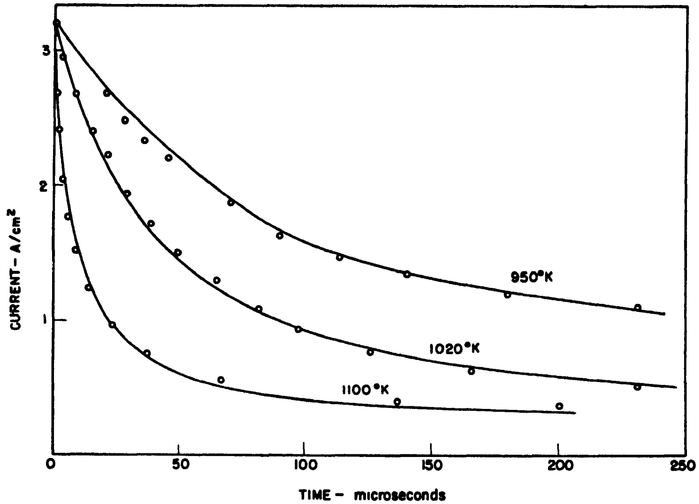


FIG. 24.—Short time emission decay at several cathode temperatures, (BaSr)O. Scales adjusted to fit theoretical equation. See ref. 62.

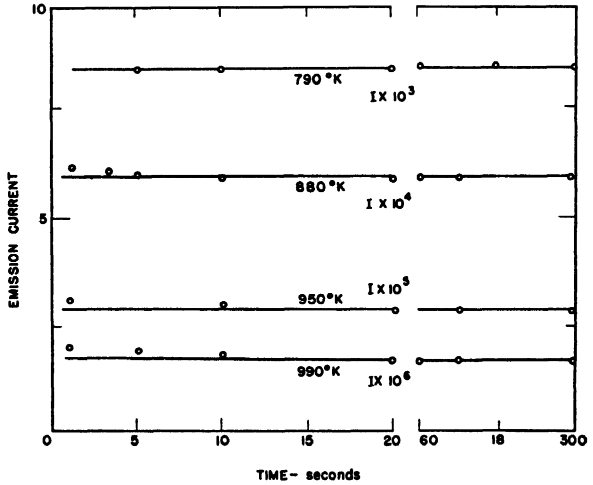


FIG. 25.—Absence of long time emission decay, BaO, compare with Fig. 23.

was limited by a decay effect which occurred at near space charge limited conditions. A similar decay, at lower emission levels, had disappeared with suitable aging of the tube, hence it was concluded that 14 $A/\text{cm.}^2$ represented only a limitation of the diode and not the ultimate d.-c. emission from the cathode. In a similar experiment Dillinger[86] very recently obtained a d.-c. emission of $18A/\text{cm.}^2$ at 1100°K.

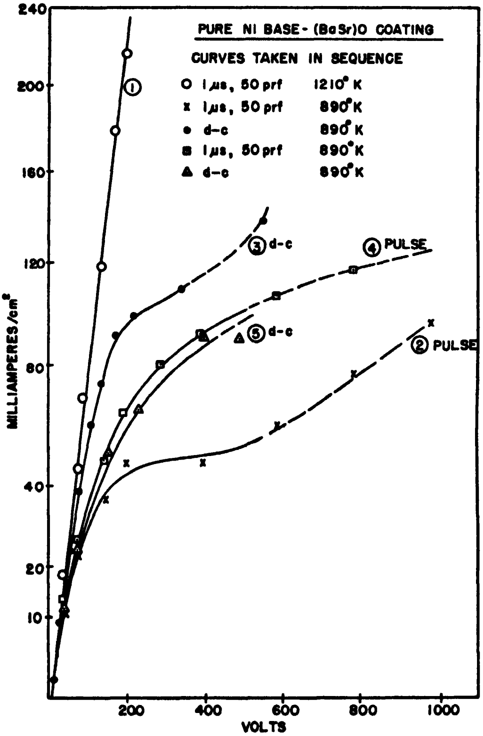


FIG. 26.—Comparison of d.-c. and microsecond pulsed emission characteristics of a (BaSr)O cathode.

Following this approach to the question of d.-c. vs. pulsed emission, the writer compared the respective emissions at a relatively low cathode temperature. Under these conditions, the saturated d.-c. emission level is sufficiently low that anode dissipation is not a serious problem. The curves shown in Fig. 26 were taken in sequence and indicate no great difference in the d.-c. and pulsed capabilities of the cathode. Time dependent emission instabilities were observed only in the dashed regions of the curves.

3. Coomes[27] suggests that a different fundamental mechanism is responsible for microsecond pulsed emission than is required for steady

state d.-c. emission. According to this, pulsed emission results from a transient phenomena, the depletion of electrons from the conduction band which is refilled between pulses by thermally excited electrons from impurity levels. A simple calculation shows that there are probably insufficient conduction electrons to supply a pulse of say 100 ampere microseconds/sq. cm. of cathode area. Thus it is necessary to presume that this theory implies an appreciable flow of current from the base metal to the coating during the pulse, induced perhaps by high electric fields established in the interface. This theory, if true, might account for the decay observed by Sproull although it is doubtful that Blewett's long time decay could be so explained.

In view of these factors, all of which require further study, there is no incentive to attempt a conclusion regarding the mechanism of decay or even the conditions under which decay exists.

9. *Sparking*

Cathode sparking or flashing, as it is referred to by the British, frequently imposes the only limitation to the emission current density which can be obtained during microsecond pulses. This phenomena is not necessarily deleterious to the cathode's operation except for an erosion of the cathode coating, for each spark is accompanied by a physical loss of coating. In magnetron operation, sparking may occur at the rate of several sparks per minute throughout the life span of the tube.

Two types of sparking are recognized; one, which produces a disruptive discharge in the tube and leaves craters penetrating throughout the whole coating and interface, and another, which appears as a scintillation on the cathode surface and leaves shallow pits in the coating. The former has been attributed to an interface breakdown,[51] and the latter may be a function of the surface oxides.[27]

Limiting our discussion to the disruptive spark phenomena, it is clearly evident that its occurrence is determined by the flow of current rather than an electric field at the cathode surface. Sparking of this type is observed under space charge limited emission conditions and the sparking current[27] decreases with increasing pulse length.

Additional evidence for placing the sparking mechanism at the interface is seen in Fig. 27 which compares the microsecond pulsed emission characteristics of cathodes prepared on a pure nickel base metal and on a 2% silicon-nickel alloy. The sparking points occur at much lower current densities in the cathode known to contain a thick, low conductivity interface layer. As the temperature is lowered, the conductivity is reduced, see Fig. 7, and likewise sparking current is seen to decrease. A second major difference in the emission characteristics of the two cathodes is the manner in which the curves deviate from the calculated Langmuir-Child space charge limited emission line. Rather abrupt breaks from this line are noted when saturation emission is reached in the pure nickel base cathode, but only a progressive deviation is seen for the silicon-nickel cathode. The point of deviation appears to be at zero current. A discussion of this phenomena will follow later.

Wright[51] assumed that the potential drop, which he observed in the coating of an emitting cathode, appeared primarily at the interface. Reasonable values of this voltage and estimates of the interface thickness led to a voltage gradient of 10^6 volts/cm. in the interface layer. Unfortunately, the probe method is not a practical tool for measuring the interface voltage at the time of sparking for this phenomena frequently "burns out" the probe leads. However, this method is useful in establishing that in certain cathodes the interface voltage is only slightly less than the total cathode voltage drop.

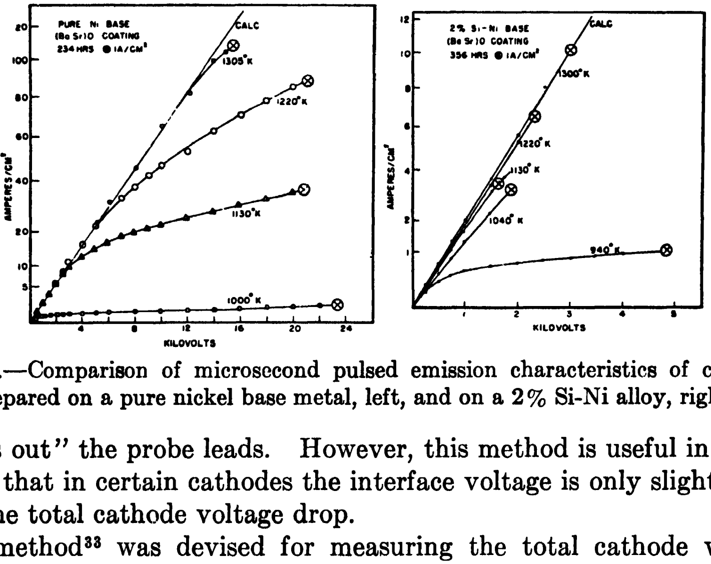
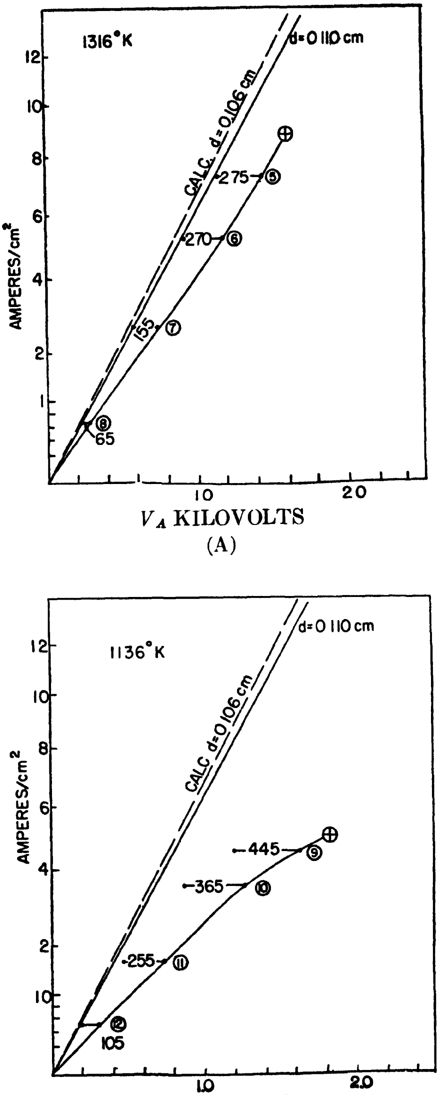


FIG. 27.—Comparison of microsecond pulsed emission characteristics of cathodes prepared on a pure nickel base metal, left, and on a 2% Si-Ni alloy, right.

A method[33] was devised for measuring the total cathode voltage drop at any point on the voltage-current characteristic. This has been useful in assigning a value to the interface voltage at near sparking conditions as well as interpreting the anomalous progressive deviation from the Langmuir-Child line. The actual voltage which appears between the external surface of the oxide coating, $V_A - V_{ic}$ in Fig. 10, can be deduced from measurements of the retarding potential required to stop electrons which pass through a small hole in the anode. A comparison of this value with the applied anode-cathode voltage allows an evaluation of V_{ic}. Typical results on a silicon-nickel base cathode are seen in Fig. 28. In (A) the experimental tube characteristic, points 5,6,7, and 8, has been corrected for the voltage drop across the cathode. These points are then seen to be in reasonable agreement with the calculated space charge



(A)

(B)

FIG. 28.—Experimental diode emission characteristics corrected for voltage drop in the interface and coating, see text.

limited emission line. The experimental points are always in better agreement with a line calculated on the assumption that the diode spacing is the distance between the outer coating surface and the anode rather than the distance between the cathode base metal and the anode. Similar conclusions were reached by Mutter for the tube geometry shown in Fig. 20; corrections for V_{ic} in both cathodes led to agreement with the coating to coating spacing. Thus it is concluded that the cathode is operating under space charge limited conditions notwithstanding the progressive deviation type of tube characteristic. In (B) the same cathode is obviously operating emission limited at points 9,10, and 11 since the correction does not return the points to the calculated space charge line. Point 9, near the sparking point, shows a total cathode drop of 445 volts of which at least 300 volts appears across the interface, see Fig. 21. An interface thickness of between 5×10^{-4} and 10^{-3} cm. for this cathode, from Table II, results in a voltage gradient of 6×10^5 to 3×10^5 volts/cm.

Such values of potential gradient clearly suggest that a breakdown phenomena initiates the cathode spark. The mechanism of this has not been discussed although the principles underlying dielectric breakdown are well established.[87]

10. Secondary Emission

Were it not for the sizeable secondary emission coefficient oxide cathodes exhibit at operating temperatures, it is questionable whether these cathodes would be able to satisfy the emission requirements of magnetron operation. In certain tubes the cathode is capable of supplying only one-tenth of the total current requirements in the form of primary emission. Values of the secondary emission coefficient, their temperature dependence, and its interpretation have been studied but remain subject to considerable controversy. Details of these investigations may be found in a separate discussion of "Secondary Emission" in this volume.

V. THIN OXIDE FILM PHENOMENA

The emission theories of Reimann and Murgoci[8] and Lowry[7] depend in common on the ability of a monolayer film of barium on the base metal to supply the entire thermionic emission of the cathode. The possibility that a film of alkaline earth oxide on the base metal, as distinguished from a film of barium, could supply this emission has been considered by G. E. Moore who has given the results below and discussed them at several scientific meetings.[88] While it is clear that such films could supply emissions of the same order of magnitude as observed in oxide cathodes, Moore does not believe that this should be taken as proof that the electrons in commercial oxide cathodes are necessarily emitted at the interface and then diffuse through the pores of the oxide.

1. *Thickness Dependence*

Sources have been developed which permit the evaporation of alkaline earth oxides without contamination from the free metal. A clean, metal receiver filament, placed along side the evaporator, is activated and its emission characteristics are determined for different amounts of adsorbed oxide. A knowledge of the rates of evaporation and an assumed size of the adsorbed molecule permit the degree of surface contamination to be expressed in units of monolayers. Fig. 29 shows typical results for the evaporation of SrO onto a molybdenum receiver. The temperature necessary to give a standard emission is plotted as a function of the amount of SrO deposited, assuming that each molecule occupies 6.6×10^{-16} cm.^{2} of surface. Many curves show no minima but bend at about 1 monolayer and slope gently downward, up to at least 20 monolayers.

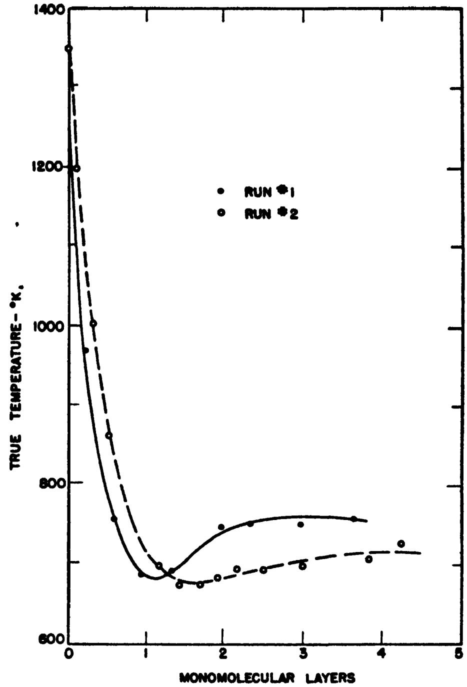


FIG. 29.—Emission properties of thin SrO films on a Mo base metal. Temperature required to give a standard emission vs. thickness in monolayers.

Emission measurements are usually limited to low temperatures to avoid a re-evaporation of the oxide. For a monolayer on the base metal under a normal oxide coating this need not be a serious limitation since evaporation from the adjacent coating could easily replenish the supply. In addition, some evidence supports the belief that films of the order of a single monolayer may adhere very tenaciously to the base metal. In Fig. 30 are shown the results of an attempt to remove a layer of BaO, initially about 20 monolayers, from a tungsten ribbon by heating at 1100°K. At various time intervals the evaporation was interrupted briefly for measuring the emission at the four lower temperatures which are shown in this figure.

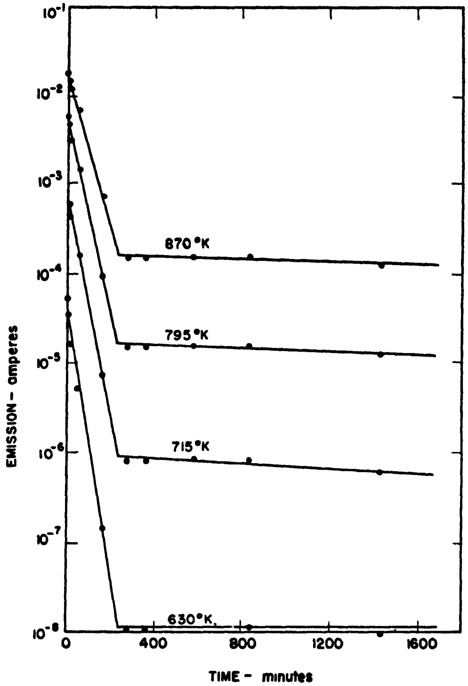


FIG. 30.—Decrease of emission accompanying removal of BaO film from tungsten by evaporation at 1100°K. Emission measured at four lower temperatures at the time intervals shown, see text.

The natural hypothesis is that the abrupt "break" in the curves at about 250 minutes of evaporation corresponds to the evaporation of all the oxide beyond the first monolayer and that the gradual decline beyond 250 minutes corresponds to the evaporation of the monolayer runs into serious difficulties. Perhaps the most serious is the requirement this would place on the vapor pressure of the oxide in the interval up to 250 minutes. Similar curves are obtained with SrO in such experiments and this hypothesis would require that the evaporation rate of SrO before the break be 10^7 times more rapid than from the crystal[1] of SrO. Other obstacles can be mentioned and the behavior is thought to represent primarily a rearrangement, rather than evaporation, although the rearrangement is not well understood. Undoubtedly, surface migration and preferential adsorption on particular crystal faces of the underlying material introduce complications which have not been considered.

2. Base Metal Dependence

Rather wide differences are found in the emission parameters when different base metals are used. Tungsten and molybdenum yield nearly identical emissions which are somewhat higher than have been obtained on nickel and much higher than the emission from platinum and gold. The explanation for these differences is not clear although it is known that only tungsten and molybdenum can be flashed readily at a temperature high enough to be assured of clean surfaces. A comparison of the emission parameters derived from Richardson plots for some of the coating-base metal combinations is seen in Table IV. Blewett's[1] value for normal thickness oxide layers and Becker's[89] values for a barium film on tungsten, are shown for comparison. It is apparent that the emission which may be taken from the oxide films is quite capable of supplying the total emission obtainable from thick oxide cathode whereas a barium film alone would hardly suffice.

TABLE IV. Emission characteristics of thin oxide film cathodes, normal thickness cathodes and metallic monolayer cathode.

Cathode system	Work function, volts	Emission, $A/\text{cm.}^2$ at 1000°K.
Thin oxide films		
CaO on W...	2.1	10^{-5}
SrO on W..	1.3	0.3
BaO on W..	1.2	0.9
SrO on Ni..	2.0	10^{-5}
Normal oxide		
CaO on Ni.	1.9	10^{-5}–10^{-3}
SrO on Ni..	1.4	10^{-4}–10^{-2}
BaO on Ni.	1.1	10^{-3}–10^{-1}
Metallic monolayer		
Ba on W..	2.1	2×10^{-3}

Having established a mechanism for emission from the base metal, our theories[7,8] must now provide a mechanism for transporting electrons through the coating to the external surface. In view of the uncertainty which now exists concerning the reflection coefficient for electrons impinging on an oxide crystal, and the ratio of ionic to electronic conductivity through the oxide crystals, a further discussion would be entirely speculative.

The success already achieved in treating the oxide cathode as an excess impurity semiconductor is most gratifying. If this success continues as more and more properties of the cathode are investigated, it may never be necessary to perform the long sought for experiment to prove which theory is correct.

ACKNOWLEDGEMENT

The author wishes to acknowledge the substantial contributions to this article made by Conyers Herring in discussing with him details of the oxide cathode theory, by G. E. Moore and W. E. Mutter for kindly permitting the use of their previously unpublished experimental results, and by the large group of researchers whose names are found in the list of references.

REFERENCES

1. Blewett, J. P. *J. Appl. Phys.*, **10,** 668–679 (1939); **10,** 831–848 (1939).
2. Wilson, H. A. *Proc. Roy. Soc. Lond.*, **A134,** 277–287 (1931).
3. Blewett, J. P. *J. Appl. Phys.*, **17,** 643–647 (1946).
4. Vick, F. A. *Science Progress*, **137,** 82–87 (1947).
5. Wright, D. A. *Nature, Lond.*, **160,** 129–130 (1947).
6. Becker, J. A. *Phys. Rev.*, **34,** 1323–1351 (1929).
7. Lowry, E. F. *Phys. Rev.*, **35,** 1367–1378 (1930).
8. Reimann, A. L. and Murgoci, R. *Phil. Mag.*, **9,** 441–464 (1930).
9. Becker, J. A. and Sears, R. W. *Phys. Rev.*, **38,** 2193–2213 (1931).
10. Jones, T. J. Thermionic Emission. Methuen & Co., London, 1936, p. 82.
11. de Boer, J. H. Electron Emission and Adsorption Phenomena. Macmillan, New York, 1935, p. 50.
12. Darbyshire, J. A. *Proc. Phys. Soc. Lond.*, **50,** 964–966 (1938).
13. Huber, H. and Wagener, S. *Z. techn. Phys.*, **23,** 1–12 (1942).
14. Seitz, F. *J. Appl. Phys.*, **16,** 553–563 (1945).
15. Maurer, R. J. *J. Appl. Phys.*, **16,** 563–570 (1945).
16. Verway, E. J. W. *Philips Tech. Rev.*, **9,** 46–53 (1947).
17. Massey, H. S. W. *J. Sci. Instrum.*, **24,** 220–224 (1947).
18. Mott, N. F. and Gurney, R. W. Electronic Processes in Ionic Crystals. Oxford University Press, 1940, p. 75.
19. Burgers, W. G. *Z. Phys.*, **80,** 352–360 (1933).
20. Benjamin, M. and Rooksby, H. P. *Phil. Mag.*, **16,** 519–525 (1933).
21. Benjamin, M. and Rooksby, H. P. *Phil. Mag.*, **15,** 810–829 (1933).
22. Veenemans, C. F. *Nederland. Tijdschr. Natuurk.*, **10,** 1 (1943).
23. Widell, E. G. *Phys. Rev.*, **69,** 247 (1946).
24. Gaertner, H. *Phil. Mag.*, **19,** 82–103 (1935).
25. Darbyshire, J. A. *Proc. Phys. Soc., Lond.*, **50,** 635–641 (1938).
26. Eisenstein, A. *J. Appl. Phys.*, **17,** 654–663 (1946).

27. Coomes, E. A. *J. Appl. Phys.*, **17**, 647–654 (1946).
28. Heinze, W. and Wagner, S. *Z. techn. Phys.*, **20**, 17–26 (1939).
29. Mecklenburg, W. *Z. Phys.* **120**, 21–30 (1942).
30. Eisenstein, A. *J. Appl. Phys.*, **17**, 434–443 (1946).
31. Seitz, F. Modern Theory of Solids. McGraw-Hill, New York, 1940, p. 188.
32. Meyer, W. and Schmidt, A. *Z. techn. Phys.*, **13**, 137–144 (1932).
33. Eisenstein, A. Journal article in preparation.
34. Benjamin, M. *Phil. Mag.*, **20**, 1–24 (1935).
35. Wise, E. M. *Proc. Inst. Radio Engrs.*, **25**, 714–752 (1937).
36. Nijboer, B. R. A. *Proc. Phys. Soc., Lond.*, **51**, 575–583 (1939).
37. Burton, J. A., private communication.
38. Mott, N. F. and Gurney, R. W. Electronic Processes in Ionic Crystals. Oxford University Press, 1940, p. 188.
39. Seitz, F. Modern Theory of Solids. McGraw-Hill, New York, 1940, p. 68.
40. Bredennikowa, T. P. *Physik. Z. Sowjetunion*, **2**, 77–90 (1932).
41. Prescott, C. H. and Morrison, J. *J. Amer. Chem. Soc.*, **60**, 3047–3053 (1938).
42. Wooten, L. A. *ASTM Bull.*, **108**, 39–44 (1941).
43. Wooten, L. A. *Phys. Rev.*, **69**, 248 (1946).
44. Mott, N. F. and Gurney, R. W. Electronic Processes in Ionic Crystals. Oxford University Press, 1940, p. 88.
45. Seitz, F. Modern Theory of Solids. McGraw-Hill, New York, 1940, p. 661.
46. Ewles, J. *Proc. Roy. Soc. Lond.*, **A167**, 34–52 (1938).
47. Arnold, H. D. *Phys. Rev.*, **16**, 70–82 (1920).
48. Rooksby, H. P. *J. R. Soc. Arts*, **88**, 318 (1940).
49. Rooksby, H. P. *G. E. C. Journal*, **11**, 83 (1940).
50. Rooksby, H. P. *Nature, Lond.*, **159**, 609–610 (1947).
51. Wright, D. A. *Proc. Roy. Soc. Lond.*, **190**, 394–417 (1947).
52. Fineman, A. and Eisenstein, A. *J. Appl. Phys.*, **17**, 663–668 (1946).
53. Hedvall, J. A. Reaktionsfahigkeit fester Stoffe. Johann Barth, Leipzig, 1938, p. 85.
54. Eskola, P. *Amer. J. Sci.*, (5) **4**, 331–375 (1922).
55. Eisenstein, A. *Phys. Rev.*, **71**, 473 A (1947).
56. O'Daniel, H. and Tscheischwili, L. *Z. Krist.*, **104**, **348** (1942).
57. Herring, C., private communication.
58. Eisenstein, A. *J. Appl. Phys.*, **17**, 874–878 (1946).
59. Schottky, W. and Rothe, H. Handbuch der Experimentalphysik. Vol. 13, Part 2, Akademische Verlagsgesellschaft, Leipzig, 1928, p. 233.
60. Schottky, W. Das freie Elektron in Physik und Technik, Springer, Berlin, 1940, p. 48.
61. Nishibori, E. and Kawamura, H. *Proc. Phys. Math. Soc. Japan*, **22**, 378–383 (1940).
62. Sproull, R. L. *Phys. Rev.*, **67**, 166–178 (1945).
63. Mott, N. F. *Proc. Roy. Soc., Lond.*, **171(A)**, 27–38 (1939).
64. Joffe, J. *Elect. Commun.*, **22**, 217–225 (1945).
65. Huxford, W. S. *Phys. Rev.*, **38**, 379–395 (1931).
66. Nishibori, E., Kawamura, H., and Hirano, K. *Proc. Phys. Math. Soc. Japan*, **23**, 37–43 (1941).
67. Heinze, W. and Hass, W. *Z. techn. Phys.*, **19**, 166 (1938).
68. Fan, H. Y. *J. Appl. Phys.*, **14**, 552–560 (1943).
69. Champieux, R. *Ann. Radioélectricité*, **1**, 208–235 (1946).
70. Heinze, W. and Wagener, S. *Z. f. Phys.*, **110**, 164–188 (1938).

71. Meyerhof, W. E. and Miller, P. H. *Rev. Sci. Instrum.*, **17**, 15–17 (1946).
72. Brown, B. B. MIT Thesis, 1942, unpublished.
73. Rose, A. *Phys. Rev.*, **49**, 838–847 (1936).
74. Morgulis, N. *J. Phys., USSR*, **11**, 67–71 (1947).
75. Seifert, R. E. and Phips, T. E. *Phys. Rev.*, **56**, 652–663 (1939).
76. Turnbull, D. and Phips, T. E. *Phys. Rev.*, **56**, 663–667 (1939).
77. Hannay, N. B. *Phys. Rev.*, **72**, 153 (1947).
78. Dillinger, J. R. Private communication.
79. Mutter, W. E. Private communication.
80. Loosjes, R. and Vink, H. J. *Philips Research Repts.*, **2**, 190–204 (1947).
81. Mutter, W. E. *Phys. Rev.*, **72**, 531 A (1947).
82. Bardeen, J. *Phys. Rev.*, **71**, 717–727 (1947).
83. Seitz, F. Modern Theory of Solids. McGraw-Hill, New York, 1940, p. 194.
84. Blewett, J. P. *Phys. Rev.*, **55**, 713–717 (1939).
85. Fineman, A. Experiments carried out at the Radiation Laboratory, MIT; private communication.
86. Dillinger, J. R. University of Wisconsin Thesis, 1947, unpublished.
87. Mott, N. F. and Gurney, R. W. Electronic Processes in Ionic Crystals. Oxford University Press, 1940, p. 197.
88. Moore, G. E. and Allison, H. W. *Phys. Rev.*, **65**, 254A (1943).
89. Becker, J. A. *Trans. Faraday Soc.*, **28**, 148–158 (1932).

Secondary Electron Emission

KENNETH G. McKAY

Bell Telephone Laboratories, Murray Hill, N. J.

CONTENTS

Page
Introduction 66
I Pure Metals 67
  1.1 Yield 67
  1.2 Shape of the δ vs V_p Curve 70
  1.3 Soft X-Ray Critical Potentials 72
  1.4 Effect of Work Function on the SE Yield 72
  1.5 Effect of Crystal Structure on the SE Yield 73
  1.6 Effect of Temperature on SE Yield 75
  1.7 Effect of Angle of Incidence of Primary Electrons on SE Yield 76
  1.8 Effect of Primary Current on SE Yield 77
  1.9 Effect of Mechanical Condition of Surface on SE Yield 77
  1.10 Effect of Adsorbed Gas on SE Yield 78
  1.11 Properties of Secondary Electrons 79
  1.12 Angular Distribution of Secondary Electrons 79
  1.13 Shot Effect 80
  1.14 Time of Liberation of Secondaries 80
  1.15 Velocity Distribution of Secondary Electrons 81
  1.16 Velocity Distribution of "True" Secondary Electrons 82
  1.17 Velocity Distribution of Reflected Primary Electrons 83
  1.18 Range of Primary and Secondary Electrons 87
  1.19 Theory of Secondary Electron Emission from Metals 90
  1.20 Rudberg and Slater's Theory 91
  1.21 Wooldridge's Theory 91
  1.22 Kadyschevitsch's Theory 92
  1.23 Conclusions about Existing Theories 93
  1.24 Methods of Measurement of SE for Metallic Targets 93
II Insulators 97
  2.1 Secondary Emission from Insulators 97
  2.2 Sticking Potentials and Yield 97
  2.3 Saturation of SE Yield 99
  2.4 Effect of Temperature and Conductivity on SE Yield 100
  2.5 Velocity Distribution of Secondary Electrons from Insulators 104
  2.6 Miscellaneous Properties of SE from Insulators 105
  2.7 Double Layer Formation and Field Enhanced Emission 106
  2.8 Theories of Secondary Emission from Insulators 109
  2.9 Methods of Measurement of SE for Insulating Targets 110

III. Composite Surfaces 114
  3.1 Composite Surfaces and Thin Film Phenomena 114
  3.2 Yield from Photocathodes of the Form [Ag]-Cs_2O, Ag-Cs 114
  3.3 Yield from Other Oxidized Targets 116
  3.4 Malter Effect (Thin Film Field Emission) 117
  Bibliography 120

INTRODUCTION

This article is a review of the present state of knowledge of secondary electron emission by electrons from metals, insulators, and complex surfaces. The technical applications of secondary emission have been fully dealt with in some of the publications listed in the bibliography and consequently have not been considered in this text. Nor has secondary emission induced by ion bombardment been included, since, theoretically, it constitutes a different phenomenon and should be considered as a separate field. For the most part, the material presented arises from work which has been performed within the past ten years. However, some aspects of the subject have not been actively investigated recently and since secondary emission, as a complete subject, has been treated rather sporadically in English scientific literature, a summary is presented of much of the earlier work. The standard reference on the subject up to 1936 is a German review by Kollath.[72] A more recent publication in book form by Bruining[216] deals with the subject up to 1941.

For convenience, the bibliography has been split into two parts: the first part consists of text references to secondary emission publications prior to 1936 and other incidental text references; the second part forms a reasonably complete list of the secondary emission publications since 1936. In cases where essentially the same material has appeared in more than one article, only one reference has been given. For a complete bibliography prior to 1936, Kollath[72] and Bruining[216] should be consulted.

It would be desirable at this stage to give an accurate, concise definition of secondary emission but this is not easy. The most general definition is that secondary emission consists of the emission of electrons from a solid due to the impact on that solid of "primary" electrons. This applies reasonably well in the case of pure metal targets. However, it breaks down for some targets of insulators where several mechanisms are invoked which are quite different from that which is normally considered to be the mechanism of secondary emission. For example, the Malter effect is accepted as field emission which is merely initiated by electron bombardment. Until more is known about the details of field enhanced emission or bombardment enhanced thermionic emission, it cannot be said whether these should be classified as true secondary emission or not. For the purposes of this article, the general definition is assumed and the following symbols or abbreviations are used:

SE: Secondary electron emission.
δ: Yield, equal to the total number of emitted electrons divided by the total number of bombarding or primary electrons.
V_p: Energy of the primary electrons, which are assumed to be monochromatic, in electron volts.
V_s: Most probable energy of emission of secondary electrons in electron volts.

I wish to express my thanks to Dr. J. B. Johnson for many stimulating discussions in the course of the preparation of this review.

I. PURE METALS

1.1 Yield

The most widely investigated property of SE is the yield as a function of primary electron energy V_p. Unless stated otherwise, it is always assumed that the primary electrons impinge normal to the surface and that all the emitted secondaries are collected. The plot of δ vs. V_p is the yield curve and has the same general shape for all materials, i.e., for low V_p, δ is much less than unity; it increases with V_p and for pure metals reaches a maximum value not greater than $\delta = 2$ for V_p equal to a few hundred volts. δ then decreases slowly as V_p is further increased. Thus although the entire yield curve may sometimes be required or the yield at some specified value of V_p, the various yield curves are sufficiently similar so that it is often adequate to specify merely the maximum value of δ, (δ_m), and the value of V_p, ($V_{p\,\max}$), for which this occurs.

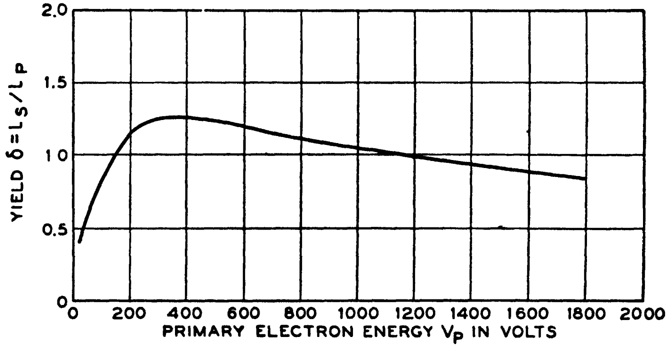


FIG. 1.—Yield curve for molybdenum.[72]

As the effects of surface contamination became apparent a great deal of effort was expended in producing cleaner surfaces which would yield more accurate values for δ. Warnecke stipulated that the metal should be given as extensive a heat treatment as possible until it reached an "end point" after which any further heat treatment made no change in the yield. Even this is open to criticism in the case of some metals such as aluminum where the oxide is much less volatile than the metal itself. Thus if the surface is once oxidized, it may not be possible to clean it off purely by heat treatment. Bruining and others have attempted to overcome this objection by using targets which have been

TABLE I. Maximum secondary emission yields of various clean metals and some semiconductors.

Element	$\delta_{\max}$	$V_{p(\max)}$ volts	Principal references
Ag	1.5	800	87, 56, 72
Al	1.0	300	66
Au	1.46	800	56, 230, 72
Ba	0.83	400	66
Be	0.6	200	66, 230
C	1.0	300	86
Cd	1.1	400	230
Co	1.2	600	152, 108
Cs	0.72	400	66, 95
Cu	1.3	600	87, 56, 230, 72
Fe	1.3	350	72, 7
K	0.7	200	195, 134
Li	0.5	85	66
Mg	0.95	300	66, 95
Mo	1.25	375	56, 72
Nb	1.2	375	56, 72
Ni	1.3	550	152, 66, 56, 72
Pd	1.3	250	5
Pt	1.6	800	72
Rb	0.9	350	216
Th	1.1	800	66
Ti	0.9	280	86
W	1.4	600	56, 121, 72
Zr	1.1	350	86
....	....	....	....
B	1.2	150	250
Ge	1.2	400	250
Si	1.1	250	250
Ag_2O	1.1	...	120
Cu_2O	1.2	...	120
MoO_2	1.1	...	120
MoS_2	1.1	...	120
SnO_2	1.1	600	258
WS_2	1.0	...	120

evaporated on in vacuum. Even where relatively clean surfaces have been obtained, the yield will still be affected by the degree of roughness of the surface and by the crystal structure. Thus it is not surprising that even the best results on a metal such as tungsten, which can be outgassed more thoroughly than most metals, exhibit considerable variation from author to author. For these reasons, the compilation of values of δ_{max} in Table I represents only what is believed to be the best available results. In most instances, for a well outgassed, smooth polycrystalline target, the values of δ_{max} are probably consistent to better than 10%. However, since many yield curves have a very broad maximum, the values of $V_{p(max)}$ may exhibit much wider variations. The principal

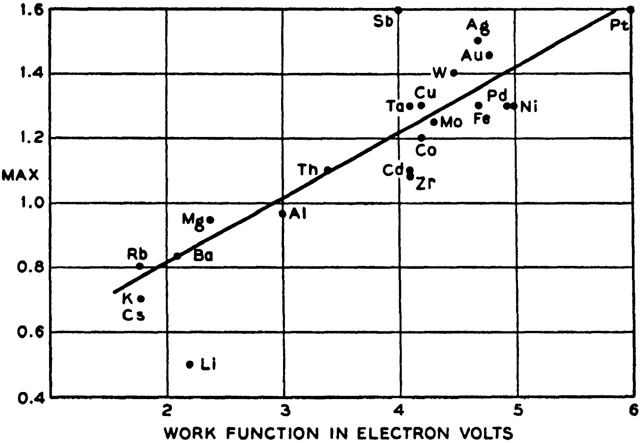


FIG. 2.—Relation between δ_{max} and work function of different targets.

references from which these data have been compiled are listed in the table and these should be consulted for the complete yield curves. In general, if the values quoted are substantially the same as that given by Kollath,[72] only that and more recent references are given. The most striking change since Kollath's compilation is the low values of δ_{max} for the low work function metals such as the alkalies. Due to outgassing problems, it is very probable that much of the early work was actually done on alkali oxides, which may give values of δ_{max} several times that of the pure metal. Data on certain semiconductors have also been included. The only justification for this is that the yields of these substances are much like those of metals. Nothing has been published concerning their other SE properties and these may not necessarily be similar to metallic behavior.

A correlation between work function and yield is illustrated in Fig. 2 where the values of δ_{max} from Table I have been plotted against work function. These values of work function have been obtained from Becker[28] and most are for polycrystalline surfaces. Here again we do not know the surface conditions, particularly the predominant crystal orientation. This and other factors render unreliable some of these work function values so that too much importance should not be attached to this plot. However, this does illustrate a possible classification of pure metals, i.e., those with high work functions have high SE yields and those with low work functions have low SE yields. The inference should not be drawn that by increasing the work function of a metal the SE yield will be increased since, as will be discussed later, this is believed to be untrue. Rather, it is believed that the work function itself plays a relatively minor role in determining the SE yield but that it is linked with other properties of the metal which play a dominant role in determining the yield. Almost as good a correlation can be obtained by plotting δ_{max} against the density of the target.

Attempts have been made to correlate δ_{max} with position in the periodic table, atomic constants, etc., but these have usually produced trends rather than correlations.

1.2 Shape of the δ vs. V_p Curve

A rough explanation for the maximum in the δ vs. V_p relation can be given by defining d_p = maximum depth of production of secondaries and d_s = maximum depth from which secondaries can escape. Then for $V_p < V_{p(max)}$, $d_p < d_s$ and $\frac{\partial \delta}{\partial V_p} > 0$. Similarly for $V_p > V_{p(max)}$, $d_p > d_s$ and $\frac{\partial \delta}{\partial V_p} < 0$. Khlebnikov[95] pointed out that according to this formulation, changes in the surface potential which affect δ but which do not affect d_p or d_s should have no effect on the value of $V_{p(max)}$. Thus it should be possible to distinguish between volume effects and solely surface effects by measurements of $V_{p(max)}$. However, since the maximum is usually rather flat, this is not a particularly sensitive criterion.

Extending this concept, Geyer[225] has plotted $\log\left(\frac{\partial \delta}{\partial V_p}\right)$ against V_p in the range 50 volts $< V_p < V_{p(max)}$, using his own experimental data and also data from Bruining and de Boer, and Copeland. He obtained linear relations in all such plots and by extrapolating to $V_p = 0$ obtained values of the zero voltage intercept $\vartheta_0 = \log\left(\frac{\partial \delta}{\partial V_p}\right)_{V_p=0}$. He found that for a given target material, ϑ_0 was independent of surface contamination, work function, angle of incidence, and depth of penetration. Moreover, metals with the same value of the principal quantum number "n" of the outermost electron shell gave the same value of ϑ_0. However, the same metals in compounds yielded much different values of ϑ_0. From this Geyer concluded that ϑ_0 is a function of the mechanism of the generation of secondaries.

Such plots depend upon careful determinations of the slope of the yield curve. Small errors in the latter may result in large errors in the slope. The author has attempted to fit data other than those given by Geyer to such plots with considerable lack of success. In particular, various published curves on tungsten by Ahearn,[20] Warnecke,[56] Coomes,[121] and McKay[221] give widely differing results, and even the values for tungsten by Sixtus[15] which are quoted by Geyer do not agree with his classification. Moreover, it is not possible in general to fit Wooldridge's[151] theoretical curves to such a linear plot. However, in spite of these contradictions, careful analyses of this type may yield information about the mechanism of secondary electron generation.

Considering the section of the yield curve where V_p is greater than $V_{p(\max)}$, Chaudri and Khan[191] have plotted log δ against V_p and obtained a linear relation for nickel. They show that if all the energy of the primary electrons is used up in the production of secondaries and if the absorption of secondaries in the metal is exponential, a relation of the

$$\delta = \delta_{\max} e^{\alpha(V - V_{p(\max)})} \tag{1}$$

form is obtained agreeing with their experimental results for $600 < V_p < 4000$ volts. To obtain this formula they assume that practically all the secondaries are produced at some point at which the primary energy has dropped down to a certain value. This is certainly an oversimplification of the process. Moreover, as V_p increases, the importance of reflected primaries also increases and this aspect was neglected. It is difficult to check this against other data in the literature since usually the yield has decreased very little from the maximum at the highest values of V_p used. However, Trump and Van de Graff[259] have measured some SE coefficients in the range 30 kv $< V_p <$ 340 kv in which they have separated the yield due to high energy secondaries or reflected primaries from the total yield. If the log $\delta/\delta_{\max}$ is plotted against V_p for the yield due to low energy secondaries from their data, a linear relation does not exist. The values of $\delta_{\max}$ were taken from Table I but small variations of these values still do not produce a linear relation. Thus apparently Chaudri and Khan's formula does not hold for very large values of V_p and certainly does not when reflected primaries are included.

Copeland[26] and Warnecke[84] have plotted the slope of the yield curve at a certain point where V_p is greater than $V_{p(\max)}$ against the atomic number of the target and have obtained a general decrease with increasing atomic number. This could best be approximated by a linear relation but the agreement was not very good. It will be seen that such a procedure is equivalent to an approximate determination of α in Chaudri's formula when the yield curve falls off very slowly.

1.3 Soft X-Ray Critical Potentials

A great deal of early work in SE was devoted to a careful analysis of the "fine structure" of the yield curve in which numerous small humps and inflexion points between zero and about 40 volts were revealed. These were correlated with the energy required to produce various soft x-rays which, according to a theory by Richardson,[18] in turn produced secondary electrons. However, later experiments showed that most of this fine structure disappeared following a really thorough degassing of the target and it is now believed to have arisen from the excitation or ionization of adsorbed gas atoms on the surface. Possibly some was due to elastically reflected primaries. Nevertheless, Warnecke[57] has observed that even after a very thorough outgassing, a few inflexion points still remain for tungsten, tantalum, and nickel. An attempted correlation between these points and possible soft x-ray energies exhibits discrepancies which would appear to be greater than the experimental error. It is possible that they might be related to the target band structure.

1.4 Effect of Work Function on the SE Yield

One method of determining the effect of the work function on the SE yield, without varying any other parameter which might affect the yield, is to deposit a thin layer of a different element on the target such that the resulting change in work function can be measured independently. De Boer and Bruining[118] have calculated that an adsorbed layer, which is equal to or less than a monomolecular layer thick, should contribute a negligible amount to the SE yield due to secondaries arising within the adsorbed layer itself, provided V_p is greater than about 50 volts. This method has been used by Sixtus[15] using thorium on tungsten, Treloar[83] with barium on tungsten and thorium on tungsten, de Boer and Bruining[118] with barium on tungsten, Coomes[121] with thorium on tungsten, and McKay[221] with sodium on tungsten. In these experiments, the work function was measured either thermionically, photoelectrically, or by contact potential. All the results, except those of Coomes, have shown a systematic increase in SE yield as the work function is lowered, passing through a maximum coincidentally with the attainment of the optimum adsorbed layer thickness for minimum work function. Both Sixtus and Treloar obtained relations of the form $\log \delta = A - b\varphi$ where φ is the surface work function and A and b are constants for a given metal and V_p. This equation, of course, holds only for adsorbed layers which are thinner than those for minimum work function. Treating the problem classically, Treloar has shown theoretical justification for such a relation and Wooldridge's[151] quantum-mechanical treatment not only predicts it but is in good numerical agreement with Treloar's experimental results. Coomes' results with thorium-coated tungsten are remarkable in that no systematic variation of yield with work function was observed although he did obtain an increase in yield with lowered work function with an oxygenated thorium coated tungsten target. The latter condition is too complex to be treated here while the former results are very difficult to explain unless the target surfaces were not what Coomes believed them to be. In this connection it is noted that Coomes' δ vs. V_p curves for clean tungsten were not completely reproducible following deposition and subsequent evaporation of a thorium layer. Afanasjeva and Timofeev[62] also obtained an increase in SE yield which passed through a maximum as potassium was evaporated on gold, silver, or platinum. However, as no corresponding maximum was observed in the photoelectric yield, this work is open to question.

It is difficult to picture any mechanism of secondary emission in which lowering of the surface work function does not increase the yield somewhat. However, what is most important is the relatively minor role that the work function plays. McKay showed that an adsorbed layer of sodium which effectively reduced the work function of tungsten by a factor of 2 increased the SE yield by only 60%, which is in rough agreement with Treloar's results of $-(\partial/\partial\varphi)(\log_e \delta_{\max}) \sim 0.12ev^{-1}$. However, this should be compared with the case of thermionic emission where such a reduction in work function would increase the thermionic emission current by a factor of about 10^6. This contrasting behavior is attributed roughly to the relatively high average emission velocity of secondary electrons to be discussed later. One consequence of this is that it appears unlikely that very high SE yields can be attained merely by a lowering of the surface work function, i.e., if Treloar's results be assumed to be valid for any value of work function, then the $\delta_{\max}$ from tungsten on reduction of the work function to zero would still be less than three.

1.5 *Effect of Crystal Structure on the SE Yield*

Work by Becker,[28] Nichols,[37] and others has established that the work function of a surface depends upon the orientation of the exposed crystal face. Thus, from §1.4 we should expect some variation in SE yield from this source. To estimate the magnitude of the effect, let us assume Treloar's result for variation of yield with work function for tungsten in conjunction with Nichols' data showing that the work function of tungsten varies from 4.35 volts for the (111) crystal direction to at least 4.65 for the (110) direction. This would give a change in δ_{max} of about 3%. However, Wooldridge's theory[151] implies that the crystal orientation may produce variations in yield other than those caused by change in work functions.

Early work by Rao[17] showed a decrease in δ_{max} from 1.3 to 0.76 in going from polycrystalline nickel to the (100) face of a single crystal of nickel. However, it is doubtful if the vacuum techniques employed were sufficient to insure that the surfaces were free of adsorbed gas.

Bekow[188] has published a preliminary note indicating that for a copper single crystal, δ is different for each crystal face and is a maximum for the (100) face. Knoll and Theile[128] have demonstrated the effect strikingly by forming an electron-optical picture of the surface of a silicon-steel target using the secondary electrons themselves. The resultant picture shows intensity variations which resemble the expected polycrystalline structure, the interpretation being that each crystal face exhibits a characteristic yield.

One method of determining the effect of crystal structure is to vary the temperature of the target and to observe changes in the SE yield as the target passes through a structure transformation point. Treloar[108] could observe no change in yield within an accuracy of a few per cent in passing through the Curie point of nickel at 358°C., the hexagonal to face-centered transformation of cobalt at 410°C., or the Curie point at 770°C. and body-centered to face-centered transformation of iron at 910°C. Wooldridge[183,184] improved the relative accuracy of such measurements to about 0.1% by adjusting V_p so that the SE yield was always unity. He observed no change in passing through the Curie point of nickel at 358°C. or of iron at 770°C. He did observe erratic changes of about 1% in the iron transformation at 910°C. but concluded that he was dealing with large crystal faces and that variations from one face to another rendered the results unreliable. He also obtained a reproducible change in SE yield of 0.4% at the hexagonal to face-centered transformation of cobalt at 410°C. Simultaneous measurements of work function by contact potential measurements showed that the change in work function was in the wrong direction to account for this change in yield.

Kollath[96,98] obtained increases in δ when evaporated beryllium targets were heated above 700°C. He attributed this to a structural change although he had no data on any known crystal transformations for beryllium at this temperature. However, the changes in δ were so large that it is doubtful if they could have been due solely to changes in crystal structure.

Suhrmann and Kundt[144,145,229] compared the SE yield of targets of copper, silver, or gold condensed at 83°K. with that obtained at room temperature. At low temperatures, the targets were assumed to be disordered and at room temperature to be ordered. They obtained yields in the ordered cases of up to 30% greater than when disordered. Due to the importance of Van der Waals' adsorption of gas at low temperatures, it would appear very likely that adsorbed gas layers might play a prominent part in this type of experiment. Wooldridge and Hartman[185] proposed that order in an alloy produces long period regularity in the lattice fields resulting in a splitting of some of the bands and making possible transitions between levels which could not exhibit interaction processes in the disordered alloy. However, measurements on a Cu_3Au target, in which disorder begins at 250°C. and is completed at 391°C., exhibited no change in SE yield greater than the experimental accuracy of 1%.

Morozov[206] has measured a change in SE yield of up to 10% in lead, antimony, and bismuth in going from the solid to the liquid state. The direction of the change depended on the bombarding velocity and also on the target. His explanation of these results is based on conductivity changes rather than crystal structure.

The effect of crystal structure on the directional scattering of reflected primary electrons will be considered in a later section.

This experimental evidence shows clearly that the crystal structure plays a role in determining the SE yield, although detailed data on the correlation between yield and structure are entirely lacking. It should be noted that even a substance which is polycrystalline does not necessarily expose all crystallite faces on the surface with equal probability. Previous cold working or heat treatments may produce preferred orientations of the surface faces. For example, tungsten wire, following heat treatment, usually develops crystals with the (110) direction parallel to the wire axis. It is possible that even before heat treatment the crystallites still maintain a preferred orientation.

1.6 *Effect of Temperature on SE Yield*

Changes in temperature of the target may alter the density of adsorbed gas if present; they may alter the crystal structure or the surface roughness. Any of these may influence the SE yield. However, when these complications are not present, many experimenters[108,121,133,221] have demonstrated no observable change in SE yield of metals with temperature. In particular Morozov[205] and Wooldridge[183,184] have shown that for cobalt, iron, molybdenum, and nickel the SE temperature coefficient must be less than the temperature coefficient of linear expansion. Such a result is in accord with a later modification of Wooldridge's theory of SE.[184]

Reichelt[169] reported an increase in the proportion of high velocity secondaries at 1500°C. with a tungsten target, resulting in an increase in the mean energy of the secondary electrons of about a volt from that at room temperature. Kollath[197] has reported measurements which contradict this and suggested that Reichelt's results might be due to an experimental error.

1.7 Effect of Angle of Incidence of Primary Electrons on SE Yield

Although most SE yield measurements have been made with normal incidence of the primary beam, a number of investigators have examined the variation of yield with incident angle. The general effect of oblique incidence may be seen by reference to Fig. 3 where for normal incidence, the average range of a primary electron is d_p. In the case of oblique incidence, secondaries produced at the end of the range d_p will be only a distance d_p cos θ from the surface, where θ is the angle of incidence, and thus have less chance of being absorbed before reaching the surface. From this we might predict that the effect will be greater for $V_p > V_{p\,\max}$ where the yield is predominantly limited by adsorption of secondaries. We might expect that for low primary velocities where the penetration is small very little variation in yield with incident angle would be observed, and this has been verified experimentally by Bruining.[41] He has also shown that a rough etched surface shows practically no variation of yield with incident angle. Such a result is to be expected since the actual incident angle is very poorly defined in this case.

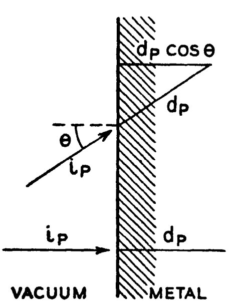


FIG. 3.—Effect of variation of angle of incidence on the path length of the secondaries.

Müller[78] investigated a series of metals with 1000 < V_p < 4000 volts and concluded that over the range $0° \leq \theta \leq 80°$, the yield varied as $(\cos \theta)^{-1}$ with deviations from this ascribed to spreading of the primary beam. However, the assumptions he used to derive a theoretical expression of this form are such that they greatly oversimplify the problem. He also showed that the relative change in yield with incident angle varied in an inverse way with the target density.

Bruining[41,88] assumed that the secondaries are absorbed exponentially with distance and derived a relation of the form

$$\delta_\theta = \delta_0 e^{\alpha x_m (1 - \cos \theta)} \tag{2}$$

where δ_θ is the yield at incidence angle $= \theta$
δ_0 is the yield at zero incidence
x_m is the mean depth of liberation of secondaries
α is the coefficient of absorption of secondaries.

This expression was verified experimentally for lithium, barium, and nickel by showing that for a given target and V_p, αx_m was a constant independent of θ within 20%. From his data it is also possible to extrapolate and obtain a value of δ_{90°, i.e., the yield to be expected when the primary beam strikes the target just at grazing angle. It might be expected that a curve of δ_{90° versus V_p would not show a maximum since there is essentially no absorption of secondaries. Actually there is a broad maximum at a mugh higher value of V_p than for normal incidence and this is attributed to the scattering of primary electrons into the target thus still producing some absorption of secondaries.

Lukjanov[99] has derived an expression for δ_θ of essentially the same form as Bruining's and has shown that it fits Müller's experimental curves. Lukjanov shows that x_m should vary inversely as the density which is in accord with the experimental fact that the greatest relative change in yield occurs with the metals of least density.

1.8 Effect of Primary Current on SE Yield

Since the SE yield is normally quoted merely as a function of V_p, it is implied that it is independent of primary current. For metals, this has been confirmed by many observers under widely differing conditions of primary current density.

1.9 Effect of Mechanical Condition of Surface on SE Yield

There is considerable experimental evidence to show that a rough or porous surface results in a lowering of the SE yield. This is qualitatively explained by postulating that a rough surface can be likened to a series of holes or wells. A secondary electron, produced at the bottom of such a well, may be trapped by the sides of the well and hence will not be emitted from the surface. Such a surface can be produced artificially by covering the target with carbon soot either smoked on or from a colloidal suspension. Optically "black" surfaces can also be prepared by evaporating various metals onto the surface through a rare gas atmosphere so that metallic agglomerates are formed before striking the target surface. Fig. 4 by Bruining[86] shows the reduction in yield with carbonized nickel. He concludes that the greatest reduction in yield occurs when the carbon granules are about 30 A. in diameter forming a fine labyrinth.[65] If other metals are used to form the surface, they should have a high sintering temperature. Otherwise, upon heating the target, the agglomerates will sinter together forming a compact high yield surface. Jonker[92] suggests a series of ribs mechanically formed on the target to reduce the SE yield still further.

1.10 Effect of Adsorbed Gas on SE Yield

In §1.4 we saw that the work function of a surface could be altered by an adsorbed metallic layer on the surface. The same, of course applies to adsorbed gas layers.[29] An adsorbed layer may also yield an appreciable number of secondaries in itself if it is sufficiently thick. However, a monatomic layer probably has a yield of the order of 0.02 for V_p around 200 volts[118] and thus the variation in yield caused by a monatomic layer is probably due almost entirely to the work function variation.

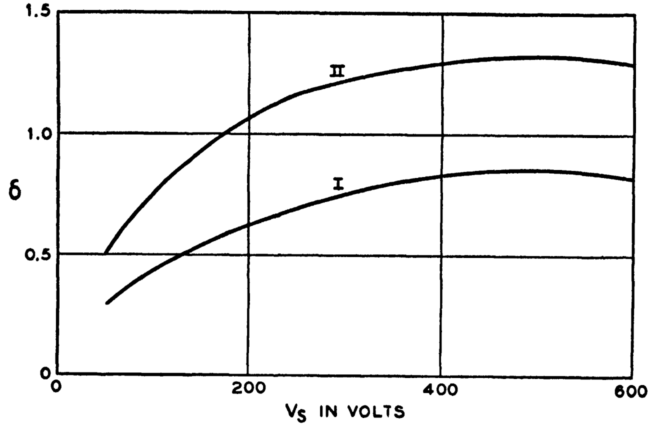


FIG. 4.—Variation of yield with surface roughness:[86] I, soot; II smooth carbon.

Most papers on SE give some data on the variation of the yield as a function of the heat treatment so the actual amount of information available is quite considerable. Of course, the heat treatment may cause changes other than the removal of adsorbed impurity layers, e.g., it may alter the crystal structure.[98] Moreover, when gas layers are being removed, they are usually of unknown thickness and composition. Thus we can only generalize that usually the removal of adsorbed gas through heat treatment causes the yield to drop possibly by as much as 50% or more and most frequently the final yield curve is lower than any other obtained during heat treatment. The removal of adsorbed gas also removes many kinks in the yield curve as described in §1.3. A good example of this general type of behavior is presented in detail by Ahearn[20] in work on the SE of tungsten. Actually the presence of oxygen on the surface may either raise or lower the yield since the physical adsorption of a monatomic layer of oxygen on most metallic surfaces causes an increase in the work function due to the formation of an electrical double layer. Such a reduction in yield has been observed.[221] However, if the oxygen forms a thick layer of oxide before or during heat treatment, the yield will be greatly altered and probably increased. Since, in some cases, the vapor pressure of the oxide is very much less than that of the metal, the oxide surface layer, once formed, can never be removed by mere heat treatment. This is probably the explanation of Warnecke's high values for the yield from aluminum.

There is but little information available about the effects of other gases. Khlebnikov[95] has increased the yield of tantalum by exposing it to hydrogen or helium. It is possible here that the hydrogen was adsorbed as ions. On the other hand, Suhrmann and Kundt[144] observed that exposure to hydrogen had no noticeable effect on the yield from copper, silver, or gold. This is in accord with the fact that physical adsorption of a monolayer of hydrogen atoms should not make any appreciable change in the work function. The behavior of complex surfaces prepared by deliberate oxidation is too extensive to be treated in this section.

1.11 Properties of Secondary Electrons

In the preceding paragraphs, the various factors which may influence the total SE yield from metals have been discussed. In the following, we shall consider some properties of the secondary electrons and related matters.

1.12 Angular Distribution of Secondary Electrons

No recent work is available on the angular distribution of secondary electrons but the results of earlier workers in the field[4,5,14] agree that if the relatively small contribution due to reflected primaries is neglected, the number of secondaries emitted per unit solid angle is greatest in the direction normal to the emitting surface and decreases with increasing angle of emergence φ as $\cos \varphi$. This is independent of the angle of incidence of the primary beam although as discussed previously the absolute value of the yield is not. A result of this type is to be expected since presumably the secondaries which are emitted at large emergence angles must, on the average, traverse a greater path length in the target than those emitted normally, and thus have a greater probability of being absorbed. It should be noted that at very high or very low values of V_p, the proportion of reflected primaries becomes appreciable and their effect is to alter the cosine distribution law as discussed later.

1.13 Shot Effect

The SE yield normally defines the average number of secondaries produced per impinging primary. This is, of course, merely a statistical mean since the total number of secondaries produced by any one primary may vary widely. The study of the resulting fluctuations can result in an estimate of the maximum time of liberation of secondaries but apart from that is now not of much importance to the theory of SE. However, the technical applications, particularly of the results of fluctuation studies by SE multipliers, are of great significance.

A number of workers such as Hayner,[31] Kurrelmeyer and Hayner,[76] Ziegler,[59,60] Shockley and Pierce,[107] and others have studied this statistical problem and their results are essentially in agreement. It would not be appropriate to derive their results here but it should be noted that analyses of the experimental results suggest that the distribution function governing the probability of emission of a given number of secondary electrons per incident primary electron, should be more general than that obtained by the use of a Poisson distribution, i.e., the relative mean square deviation of δ may not be equal to the reciprocal of δ. Shockley and Pierce have shown that for an electron multiplier with n stages each of average gain m, the mean square noise current in a frequency band Δf in the output circuit is given by

$$I_s{}^2\Delta f = M^2 I_p{}^2 \Delta f + d^2 \frac{M(M-1)}{m(m-1)} \cdot 2e\bar{I}_p\Delta f \qquad (3)$$

where $I_p{}^2\Delta f$ = mean square noise in the primary current in a frequency band Δf

$\bar{I}_p$ = average component of primary current

d^2 = mean square deviation of m

$M = m^n$ = average overall gain

e = electronic charge.

This result applies for frequencies so low that the period is large compared with the time of collection on the final anode of all the electrons descended from the same primary electron at the input, i.e., the burst duration. Sard[254] has extended the frequency range of the analysis and has shown that for the R.C.A. type 931 photomultiplier tube, the noise spectrum should be essentially uniform from zero frequency to about 100 Mc and should fall rapidly to a very low value for higher frequencies.

1.14 Time of Liberation of Secondaries

In general it has been assumed that the time between the arrival of a primary electron and the emission of a secondary is essentially zero.

By means of shot effect measurements, Hayner[81] concluded that the emission time must be less than 10^{-6} seconds and is possibly less than 10^{-8} seconds. Wang[247] has replaced the reflector in a reflex klystron by a SE surface and has obtained satisfactory operation at 4000 Mc. From this he concluded that the emission time lag must be less than 2×10^{-10} seconds or, if greater, the time lag dispersion must be less than 2×10^{-10} seconds. Greenblatt and Miller[256] made similar observations in a 3000 Mc secondary electron multiplier from which they concluded that at least some of the secondary electrons were emitted in less than 5×10^{-11} seconds. Although these measurements are all somewhat indirect, it is probable that the actual time lag in emission is less than about 10^{-12} seconds and is thus negligible in most experimental work involving SE from metals.

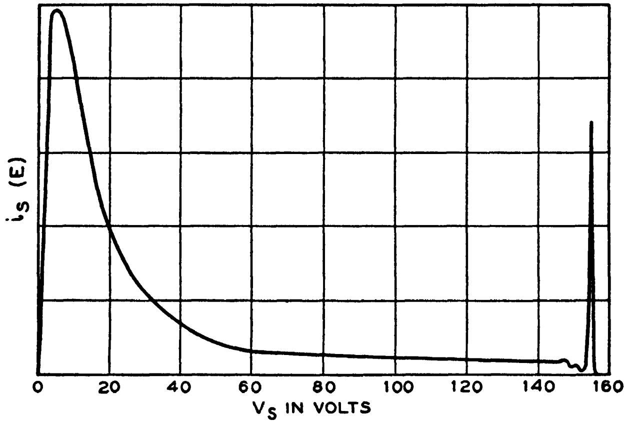


FIG. 5.—Secondary electron velocity distribution from gold.[50]

1.15 *Velocity Distribution of Secondary Electrons*

It is evident that the velocity distribution of the emitted secondary electrons is of the greatest importance, both in applications of secondary emission and in the theoretical interpretation of the process. It is strange that very little work has been reported concerning this phase recently and consequently we must rely mainly on earlier experimental work, some of which is difficult to interpret.

The general form of the velocity distribution curve resulting from medium values of $V_p (20 < V_p < 1000$ volts) has been nicely demonstrated by Rudberg,[50] one of whose curves is reproduced in Fig. 5. He used a transverse magnetic analyzer with angles of incidence and emergence both equal to 45°. The curve is interpreted as follows: the large majority of the electrons are emitted with energies of a few volts. These are regarded as true secondaries. Merged in this group and extending somewhat uniformly out to energies almost equal to V_p is a relatively small number of inelastically reflected primaries. Finally there is a sharp peak, at energy equal to V_p, of elastically reflected primaries. Apart from experiments expressly designed to investigate reflected primaries, practically all the work on secondary emission deals essentially with the "true" secondaries, i.e., low velocity secondaries. It will later be shown that complete neglect of the reflected primaries is not always possible without introducing considerable error.

1.16 Velocity Distribution of "True" Secondary Electrons

Although many measurements have been made of this aspect of secondary emission, most of them are subject to the same criticisms that have been made of the yield measurements: inadequate outgassing, absence of data on the condition of the target surface or on the crystal structure. It is very difficult to estimate the extent to which the data are affected by these parameters since but few measurements have been made of the velocity distribution as a function of anything but V_p.

The distribution curve is similar, although it does not correspond exactly, to a Maxwellian distribution. The most probable emission energy will be denoted by $V_{s\,\text{max}}$. This is always somewhat smaller than the average emission energy. According to Becker,[3] Brinsmade,[9] and others, for any given target $V_{s\,\text{max}}$ is independent of V_p for 20 volts $< V_p <$ 1000 volts in agreement with a theory by Kadyschevitsch[242] in which it is also shown that $V_{s\,\text{max}}$ should decrease as the surface work function is decreased. Such behavior has been verified by Bronstein[224] who measured the energy distribution as a function of layer thickness of silver on a nickel base.

As discussed in §1.6 there is unconfirmed evidence that the mean energy of secondaries from tungsten can be increased about 1 volt by heating the target to 1500°C.

Kushnir and Frumin[199] have reported that for molybdenum and silver $V_{s\,\text{max}}$ increases as the angle of emergence increases. There is also evidence[33,50] that the velocity distribution is affected by adsorbed gas on the target. Thus the situation is analogous to that for the yield curve: the velocity distribution is determined by a large number of parameters not all of which are known for any given experiment. Even in cases where all the most important parameters have been determined, it is not possible to compare directly the results of different experimenters since the experimental conditions differ and there is not enough known about the effect of the various parameters to enable one to extrapolate from one set of conditions to another.

Table II shows a summary of the most recent data on the energy distribution of slow secondary electrons. The methods used, as described in §1.24 are as follows: RE, retarding electric field; TM, transverse magnetic field; LM, longitudinal magnetic field. The incompleteness of the available data is evident. This is due in most part to the experimental problem of making such measurements with reasonable accuracy. Experimentally and theoretically, the problem of the energy distribution is much more difficult than that of the yield curve. Apart from Kadyschevitsch's[242] work, the problem has not been treated theoretically. Attempts to correlate $V_{s\ \mathrm{max}}$ with the atomic number of the target have not proved successful. Haworth[30,46] has detected some slight subsidiary peaks in the energy distribution curves for molybdenum and columbium but these results have not been confirmed by other experimenters, e.g., Kollath's[163] recent work on molybdenum.

TABLE II. Summary of available data on most probable secondary electron energy.

Target	Method	Inc. angle	Emer. angle	V_p volts	$V_{s\ \mathrm{max}}$ volts	Author	Reference	Year
Ag	TM	45°	45°	155	5.4	Rudberg	50	1936
Ag	RE	70°	0°	10–100	2–3	Langewalter	33	1935
Al	TM	45°	45°	36–176	5–6	Brinsmade	9	1927
Au	TM	45°	45°	155	5.4	Rudberg	50	1936
Be	LM	0°	~30°		3	Kollath	163	1940
Cb	TM	45°	45°	147	4.5	Haworth	46	1936
Cu	TM	45°	45°	155	3.3	Rudberg	50	1936
Fe	RE	30°	0°	24–2075	2	Becker	3	1925
Mo	TM	45°	45°	150	3	Haworth	30	1935
Mo	TM	45°	45°	7.5–100	4	Soller	19	1930
Mo after heat treatment	TM	45°	45°	7.5–100	10–20	Soller	19	1930
Mo	LM	0°	~30°		2.9	Kollath	163	1940
Pd	RE	70°	0°	10–100	2	Langewalter	33	1935
Pt	RE	70°	0°	10–100	2–3	Langewalter	33	1935
Ta	LM	0°	~30°		2	Kollath	163	1940

1.17 Velocity Distribution of Reflected Primary Electrons

Owing in part to the relatively small number of reflected primaries for medium values of V_p, i.e., 20 volts $< V_p <$ 1000 volts, in comparison with what here has been defined as "true" secondaries, the general behavior of reflected primaries has not received much attention. Here it is necessary to specify clearly the range of V_p with which one has to deal.

a. V_p less than about 10 volts. For very low values of V_p it has been clearly demonstrated, notably by Gimpel and Richardson[226] (for $V_p = 1$ volt), that all the secondaries are emitted with the same energy as the primaries, i.e., we are dealing with elastically reflected primaries. In such experiments, therefore, the secondary emission coefficient is replaced by a "reflection coefficient." For such low values of V_p, the effects of contact potential, inhomogeneity in the primary velocities, and adsorbed gas play such an important role as to make the experiments very difficult to perform. Gimpel and Richardson[226] showed that for a copper target, the reflection coefficient is equal to 0.24 which remains constant within 20% for $0.35 \text{ volts} \leq V_p \leq 10 \text{ volts}$. It is probable that their results were to some extent affected by adsorbed gas on the surface. Bruining[89] has also measured the reflection coefficient at low velocities and showed that for silver $\delta_{\text{reflec.}} = 0.1$ at $V_p = 25$ volts rising to 0.2 at $V_p = 3$ volts. Over the same range of V_p, barium has a $\delta_{\text{reflec.}} = 0.05$ rising to 0.1 for $V_p = 3$ volts. Over this voltage range, Bruining had to separate out the reflected primaries from the true secondaries or inelastically reflected primaries by plots of the energy distribution. The value of V_p at which the emission contains only reflected primaries is not sharply defined. Data on this are so meager that one can only say that the ratio of elastically reflected primaries to the total emission decreases from 100% for very low V_p to a very small fraction around $V_{p\ \text{max}}$.

Modern theoretical treatments[35] of the reflection of slow electrons by an image force type of surface barrier indicate that as $V_p \rightarrow 0$ the reflection coefficient should approach a limiting value of something less than 10% for most metals.

b. V_p greater than 10 volts. The type of energy distribution curve obtained in this voltage range was shown in Fig. 5. Rudberg[50] investigated the fine structure for energies approaching the primary energy, for targets of copper, gold, and silver and for composite targets of calcium on silver, calcium oxide on silver, and barium oxide on silver. He found several discrete peaks differing from V_p by a few volts. Although their magnitude varied with V_p, their position relative to V_p did not, indicating that they represent discrete energy losses due to inelastic collisions of the primary electrons. These peaks were found to be characteristic of the target material and, indeed, by evaporating barium or calcium on the target, he found that the resulting changes in the loss peak structure were sufficient to identify surface layers of only a few atom diameters in thickness. Rudberg and Slater[51] developed a quantum mechanical theory dealing with this phenomenon which will be discussed later.

Farnsworth[21,45,110] has made extensive investigations of the scattering of the elastically reflected primaries using polycrystalline or single crystal targets along the lines of the classic experiments of Davisson and Germer.[10] He has shown that for $V_p = 200$ volts, at least 90% of the elastically reflected primaries arise from the first two atomic layers. Such electrons do not obey the cosine law discussed in §1.12 but are emitted preferentially in accordance with the Bragg law. However, such studies form a rather specialized field. Although they enable us to obtain considerable information about some aspects of the interaction of electrons with matter, they are not of general interest in secondary emission since for all but very small values of V_p, these electrons usually form an extremely small fraction of the total emission. There are some special cases in which the elastically reflected primaries cannot be ignored. Davisson and Germer's[10] experiments on electron reflection showed that when a single crystal is bombarded by primaries of the right velocity and angle of incidence to satisfy the conditions for Bragg reflection, as many as 40% or more of the incident primaries may be elastically reflected. Since these primaries are reflected in beams whose directions are determined by the crystal structure, they may produce appreciable deviations from the cosine law of angular distribution of secondaries.

Although usually the elastically reflected primaries form a negligible part of the total emission, the sum of these and the inelastically scattered primaries certainly may not. Unfortunately there is very little information available on this phase. Although it is impossible to separate slow inelastically reflected primaries from the mass of true secondaries, it is perhaps useful to pick an arbitrary value of energy and, merely for the sake of classification, define as scattered primaries, those electrons which are emitted with energies greater than this value. From Rudberg's curves it is reasonable to pick 50 ev as the arbitrary distinguishing energy. Such a classification could only apply for values of V_p considerably larger than this limit. Let us assume that it holds for all V_p greater than 100 volts. Although this definition is without physical basis, nevertheless it is a very useful concept practically. For example, early work by Farnsworth[5] showed that for copper with $V_p = 102$ volts, 25% of the emitted electrons had energies greater than 50 ev. A crude integration of Rudberg's published curves shows that with $V_p = 155$ volts, the corresponding percentage of scattered primaries of over 50 volts energy is: gold 21%, silver 31%, copper 35%. Although these figures are not very exact, they are such as to indicate that the high speed scattered primaries are by no means negligible. Most SE yields are obtained with an electron gun bombarding a target which is more or less surrounded by a collector. The primary current is often measured by putting the collector at a negative potential of about 50 volts with respect to the target and measuring the target current. Obviously, this may give an erroneous value for the primary current. If these high speed electrons strike the collector, they may in turn produce secondaries which will be collected on the target thus possibly reducing the actual error but not the uncertainty. It is true that the beam current can be determined by measuring the total current leaving the cathode if it can be guaranteed that practically none of the electrons leaving the cathode strike anything other than the target. However, this is frequently not the case. Thus it must be emphasized that a detailed knowledge of the behavior of high velocity scattered primaries must be obtained before very accurate SE measurements can be assured.

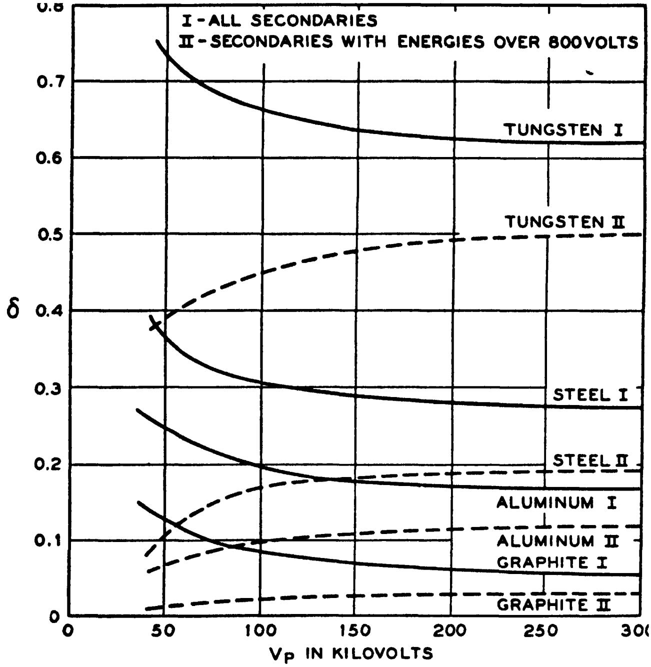


FIG. 6.—Variation of yield for high V_p.[259]

For higher values of V_p, the proportion of high velocity electrons increases. Stehberger[12] gives the following values for a gold target:

V_p Volts	Electrons Emitted with Energies Greater Than 50 Ev (in %)
1000	20
2000	28
7600	42
11000	48

Recent work by Trump and Van de Graff[259] is shown in Fig. 6. As they observed practically no electrons with energies between 20 ev and 800 ev, the dotted curves represent essentially what we have defined as high speed scattered primaries. In this work V_p is very large compared with 800 volts so the relative lack of electrons in this range is to be expected if the high speed electrons have a more or less equal probability of being obtained at any voltage between 50 volts and V_p. In Fig. 7 is shown a replot of Fig. 6 giving the percentage of scattered primaries as a function of V_p.

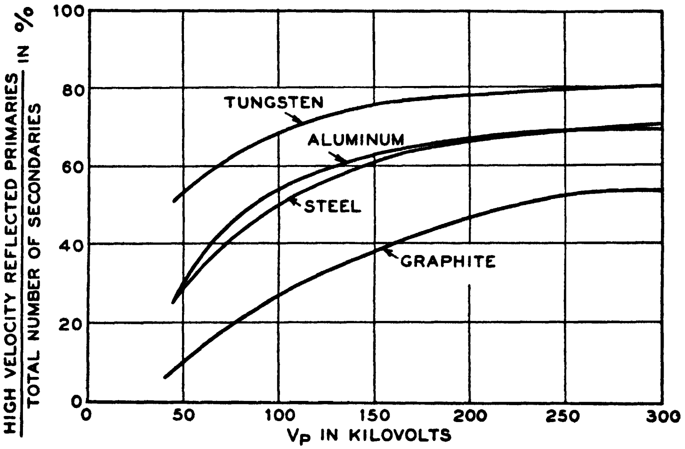


FIG. 7.—Replot of Fig. 6 showing percentage of reflected primaries as a function of V_p.

1.18 Range of Primary and Secondary Electrons

From the discussion in §1.2 on the shape of the yield curve, it is evident that two extremely important factors are: (1) the rate of loss of energy of the primary electrons and (2) the rate of absorption of the secondary electrons. The entire problem of the rate of energy loss of electrons in passing through matter is too extensive and complex to be dealt with adequately here. Rather, we shall discuss only those aspects which are pertinent to secondary emission. Bethe[16] has given an excellent wave mechanical treatment of the rate of energy loss of electrons with energies greater than several thousand electron volts, and this theory is in good agreement with experimental results. However, Bethe's equations are not valid for lower electron energies and thus cannot be integrated to give the total range, i.e., the total distance that is traversed before the electron becomes indistinguishable from an electron with thermal velocity. Moreover, the most direct experimental approach, which consists of measuring the electron transmission of thin foils, becomes very difficult for low velocity electrons. The foils must be extremely thin and there is considerable danger that most of the transmission is due to thin spots or actual holes through the foil. Another potential source of error has been pointed out by Katz[70] and Was and Tol,[38] who have shown that changes in the crystal structure of such thin films can be brought about by electron bombardment which alters the transmission. Thus it cannot always be assumed that the behavior of electrons in a thin foil is the same as in the bulk material.

It is generally assumed that the electron density I decreases exponentially with distance x, i.e., $I = I_0e^{-\alpha x}$ where α is the absorption coefficient. Becker[13] has investigated the transmission of nickel foils in the low velocity region giving a value of $\alpha = 1.5 \times 10^6$ cm.$^{-1}$ which is nearly independent of primary velocity up to 1000 volts.

From classical theory, Whiddington[1] has shown that the rate of energy loss is given by

$$[eV(x)]^2 = (eV_p)^2 - ax \tag{4}$$

where $eV(x)$ is the electron energy at distance x from the surface
eV_p is the electron energy at the surface
a is a constant.

Terrill[2] has shown that for many metals $a/\rho = 0.40 \times 10^{12}$ volt^2 cm.$^{-1}$ within an accuracy of about $\pm 10\%$ where ρ is the density. As will be shown later, Bruining[216] and others have constructed classical theories of secondary emission using these two formulas. Bruining has shown that

$$eV_{p\,\max} = 0.92\sqrt{a/\alpha} \tag{5}$$

If known values of $V_{p\,\max}$ and Terrill's values of a are used, values for α can be calculated. By this means, Bruining obtained values for α ranging from 4×10^6 cm.$^{-1}$ for cesium to 2×10^7 cm.$^{-1}$ for molybdenum. The value for nickel was 1.2×10^7 cm.$^{-1}$ which is considerably in excess of Becker's experimentally determined value. This probably reflects mainly the extent of the error involved in assuming eq. (4) for the rate of energy loss of the primary electrons.

With a value for α and also data on the variation of yield with incident angle, it is possible to arrive at a figure for the mean depth of origin of the secondary electrons. As shown in §1.7 Bruining[41] obtained a value for αx_m by this method. Using Becker's value for α for nickel, he obtained for the mean depth of origin in nickel, $x_m = 30$ A. or 14 atomic layers, for $V_p = 500$ volts.

Another method of determining x_m is to cover the target with successively thicker layers of another metal with a considerably different value of δ. For very thin films δ will be characteristic of the target metal; for thick films it will be characteristic of the layer material. The intermediate region is of such a layer thickness as would correspond to x_m. Kadyschevitsch[243] has pointed out that when the SE is also a function of the work function, such a procedure may yield misleading results. Essentially the argument is that the work function of a thin layer is characteristic of neither the target material nor of the layer material in bulk. This argument would appear to be valid mainly for low velocity primaries where x_m is only a few atomic layers. Using the layer technique Copeland[155] obtained the maximum depth of origin of secondary electrons from platinum on aluminum, as shown in Fig. 8. It will be seen that the maximum depth of origin increases linearly with V_p for small V_p.

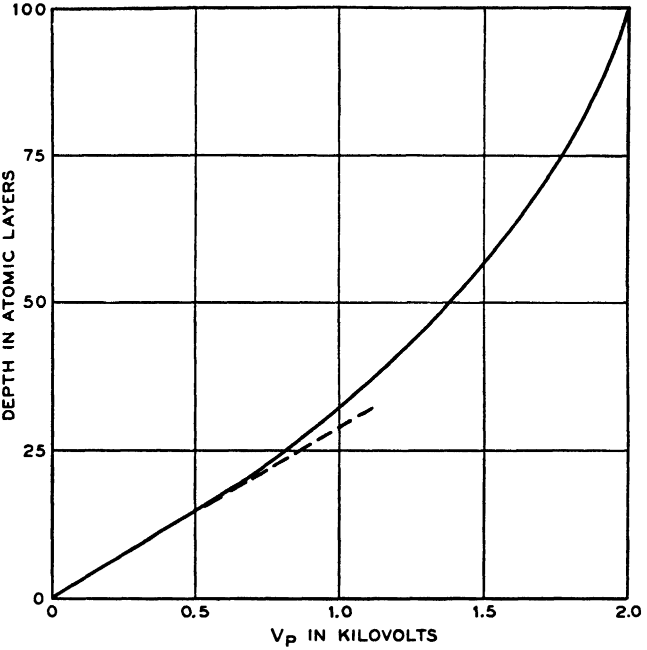


FIG. 8.—Maximum depth of origin of secondaries in platinum.[155]

Hastings,[158] studying silver on platinum, concluded that secondaries with less than 20 ev energy originate within 15 atomic layers of the surface while those of less than 50 ev energy originate within the first 30 atomic layers. He also concluded that most of the high velocity scattered primaries arise within the first atomic layer. This last conclusion is in agreement with work by Farnsworth[45] and Rudberg.[50]

Truell[228] studied layers of magnesium on carbon with primary energies from 2000 to 8000 volts. He concluded that the depth of origin of secondary electrons in the energy range from 10 volts to 200 volts varies from 2×10^{-7} cm. to 4×10^{-5} cm.

1.19 Theory of Secondary Electron Emission from Metals

A number of attempts have been made to construct a purely classical empirical theory of secondary emission.[63,171] Here we shall follow Bruining's exposition.[216] Let us assume that the primaries lose energy according to the Whiddington law (see §1.18), i.e., $[eV(x)]^2 = [eV_p]^2 - ax$. We assume that the rate of energy loss of primaries is proportional to the number of secondaries produced per unit path length and also that the secondaries are absorbed exponentially with an absorption coefficient α. Then the number of secondaries available in vacuum which arise from a layer of thickness dx at a distance x from the surface is:

$$di_s = -Ki_p e^{-ax} \cdot \frac{d[eV(x)]}{dx}\, dx \tag{6}$$

where K is the constant of proportionality. Using the Whiddington law

$$di_s = \tfrac{1}{2} K a i_p e^{-\alpha x} (e^2 V_p^2 - ax)^{-\frac{1}{2}} dx \tag{7}$$

The maximum penetration of the primaries is given by putting $eV(x) = 0$ in eq. 4. Hence $x_{\max} = \dfrac{e^2 V_p^2}{a}$. Thus

$$i_s = \frac{1}{2} K a i_p \int_0^{x_{\max}} [e^2 V_p^2 - ax]^{-\frac{1}{2}} e^{-\alpha x} dx \tag{8}$$

$$= K i_p \sqrt{\frac{a}{\alpha}}\, e^{-r^2} \int_0^r e^{y^2} dy \tag{9}$$

where $r = eV_p \sqrt{\dfrac{a}{\alpha}}$. The substitution here is $x = \dfrac{e^2 V_p^2 - (a/\alpha) y^2}{a}$. To obtain $V_{p\,\max}$ we differentiate eq. 9 with respect to eV_p and equate to zero.

$$\frac{di_s}{d(eV_p)} = K i_p \left(1 - 2re^{-r^2} \int_0^r e^{y^2} dy\right) = 0 \tag{10}$$

which is solved for $r = eV_p \sqrt{\alpha/a} = 0.92$ or

$$eV_{p\,\max} = 0.92 \sqrt{a/\alpha} \tag{5}$$

Using known values of a and α for nickel, Bruining obtained $V_{p\,\max} = 1420$ volts as against the experimental value of 550 volts.

For $V_p \gg V_{p\,\max}$ the integral in eq. 9 can be expanded by partial integration and all terms but the first neglected yielding an error of only

2% for $r = 5$. This gives

$$i_s = Ki_p \frac{a}{2\alpha \cdot eV_p} \tag{11}$$

For $V_p \ll V_{p\,\max}$, eq. 9 reduces to

$$i_s = Ki_peV_p \tag{12}$$

which is independent of a and α, i.e., there is negligible penetration of the target. It must be emphasized that these expressions do not include reflected primaries, which for the two limiting cases, become quite important.

Such a theory as this demonstrates the three main processes involved in secondary emission: loss of energy by the primaries, transfer of energy to secondary electrons, and absorption of secondaries before emission. However, nothing is said about the detailed mechanism involved in any of these processes. Inasmuch as this involves the interaction of electrons with densely packed matter, little progress was made theoretically until the advent of quantum mechanics. Since then several theories have been developed. However, a detailed exposition of these would be far too lengthy to be warranted in this review; only the principal features of each can be mentioned.

1.20 Rudberg and Slater's Theory

Rudberg and Slater[51] developed a quantum mechanical theory to account for the reflected primaries which have suffered discrete energy losses as discussed in §1.17. Using the wave functions for an infinite crystal, they considered the probability of excitation of a bound electron to a higher energy level. The inelastically reflected primary electron would demonstrate this energy loss. Detailed calculations for copper showed good agreement with experiment for small energy losses. The departures from experiment for higher losses were attributed to the fact that the free-electronlike wave functions used were not a proper approximation to the actual functions in the region of the surface.

1.21 Wooldridge's Theory

Following the work of Born, Bethe, and others, Fröhlich[22] developed a theory of secondary emission based on wave mechanics. The theory predicts about the right order of magnitude for the secondary emission and also the approximate course of the δ vs. V_p curve. Later Wooldridge[151,184] pointed out some errors in Fröhlich's treatment and formulated a more complete and quantitative theory.

Wooldridge uses Bloch eigenfunctions for the simple crystal lattice and then treats the effect of the primary electron as a perturbation problem. He shows that the principal source of energy loss by primary electrons with energy not very much greater than $V_{p\ \max}$ is caused by the production of secondary electrons which arise from loosely bound valence electrons. He shows that the contribution by inner shell electrons is very small. This differs from Bethe's theory of rate of energy loss by electrons in which he considers primaries of much higher velocity. The boundary between the two theories lies in the region of 1000 ev primary energy. Wooldridge shows that the primaries lose energy in discrete units which, in the case of silver, amount to about 25 ev. Thus primary electrons of energy lower than this value (inside the metal) should not produce appreciable SE.

Having developed an expression for the rate of loss of energy of the primaries and for the production of secondaries, he considers how many of these secondaries can escape assuming exponential absorption. Due to uncertainty concerning the lattice fields and the absorption coefficient, the final expression for the yield involves an undetermined parameter in addition to the energy constants of the metal. This parameter is determined by matching the value of $\delta_{\max}$ to that determined experimentally. By this procedure he obtains good agreement with experimentally determined yield curves for high density metals. However, the agreement is not nearly as good for low density materials. He suggests that this may be due to neglect of energy loss by the primaries due to free electron or Rutherford scattering which becomes increasingly important as the atomic volume increases. Wooldridge also estimates that the change in yield with work function should agree with experimental results by Treloar and others. (See §1.4.)

1.22 Kadyschevitsch's Theory

Kadyschevitsch[160,196,242] has developed a theory of secondary emission on somewhat different lines from Wooldridge. It is essentially based on classical dynamics and quantum mechanics is used to determine the limits of validity of such approximations and to set up the electron band picture. He justifies the use of classical mechanics by postulating a sufficiently high relative velocity between the primary and the incipient secondary electrons. This restricts the theory to the range of $V_p \gtrsim 200$ volts. He considers only the reaction between a primary electron and a perfectly free electron in the Fermi gas. He concludes that the number of bound electrons which can be emitted as secondaries is negligibly small. This is in disagreement with Wooldridge's conclusions. Since lattice interaction in the collision process is neglected, a normally directed primary cannot produce any secondaries in the metal with velocity components pointed toward the surface. Therefore he then considers the dispersion and absorption of secondaries in detail and develops expressions for the dispersion following multiple elastic collisions. His final expression for the yield contains the electrical constants of the metal (work function, etc.) and also the mean free paths for elastic and inelastic collisions of primary and secondary electrons. Since the mean free paths were not subject to direct measurement, these quantities were estimated indirectly. He concludes that when the ratio of the effective mean free paths of the primary to the secondary electrons is equal to 0.56, the yield is a maximum for normal primary incidence. For silver and nickel he obtains good agreement with experimental data for δ/V_p restricting V_p to the range 200–1400 volts, although the details of calculation are not given.

The theory is also applied to compute the variation of δ with incident angle of the primaries, the distribution of the secondary electrons in energy and direction, and the forward yield of secondaries from a thin target, all of which are in reasonable agreement with experiment.

Unfortunately, this theory contains a number of parameters which can be determined only by indirect methods. It is most difficult to estimate the extent of the possible errors involved. Consequently, data are lacking to enable one to predict the yield curve for other metals.

1.23 Conclusions about Existing Theories

At the present time, the available theories of secondary electron emission can explain the process in general terms but lack the detailed data necessary to give a complete quantitative description. As yet, little attention has been given to the effect of the surface except to postulate a surface barrier. Yet experimentally it has been demonstrated that the condition of the surface plays a considerable role in determining the yield and thus it would appear that any complete theory must consider the surface of the metal in detail.

1.24 Methods of Measurement of SE for Metallic Targets

Apart from the original discovery of secondary emission, probably the most important development has been the realization of the extensive degassing treatment and careful high vacuum techniques which are essential to give reproducible and reliable results. The importance of this was emphasized by Warnecke and by many subsequent experimenters. The subject of high vacuum technique is too extensive to be treated here and reference should be made to some of the recent papers where careful descriptions of the experimental techniques have been made.

Nearly all yield measurements on metals have been made by one of two methods, the triode method or the electron gun method, which are as follows:

In its simplest form, the triode method involves an ordinary triode tube in which the grid is positive in potential with respect to the plate, which is in turn positive with respect to the cathode. The plate forms the secondary emitter, the secondary electrons being collected by the grid. Let the current leaving the cathode be i_k of which the grid intercepts a fraction s. Then the primary current striking the plate is $i_k(1 - s)$ and the secondary current leaving the plate which is collected by the grid is $\delta i_k(1 - s)$ where δ is the secondary yield of the plate at V_p = plate to cathode potential. Thus the current in the plate circuit is

$$i_p = -(\delta - 1)(1 - s)i_k \tag{13}$$

and the current in the grid circuit is

$$i_g = \delta i_k(1 - s) + s i_k \tag{14}$$

Solving for δ we get

$$\delta = 1 - \frac{i_p}{(1 - s)(i_g + i_p)} \tag{15}$$

together with the obvious relation

$$i_k = i_g + i_p \tag{16}$$

Examination of eq. 15 shows that the fraction s intercepted by the grid must be determined independently and all the variations of the triode method hinge on this determination. These will be mentioned briefly; they are discussed in much greater detail by Treloar.[109] Several depend on the fact that for a given geometry, s is a function only of the ratio V_p/V_g where space charge is neglected.

The most direct approach is to calculate s from the geometry.[8] This requires a specially designed tube and even then is none too satisfactory. Hyatt[11] measured s by using positive ions instead of electrons with $|V_p|$ less than 100 volts where the number of secondaries produced is very small. However, he found s essentially independent of V_p/V_g which is not normally true. Lange[6] used low velocity electrons ($V_p < 10$ volts) and assumed that the secondary emission was negligible. Since we expect $\delta \sim 0.2$ due to reflected primaries, this is of doubtful validity. Moreover, when using very low voltages, the problem is further complicated by the potential drop across the filament, the initial velocity distribution of the primaries, and the effect of contact potentials and space charge. Lange also used another method consisting of a magnetic field parallel to the tube axis to suppress the relatively slow moving secondary electrons. It was shown that this is valid only for $V_p/V_g \sim 1$ and thus is of limited application. Myers[79] made use of the difference in energy of the primary and secondary electrons and measured the temperature rise of the anode. Correlating this with the secondary electron velocity distribution, he obtained a measure of s which was independent of space charge. However, the experimental techniques required are very difficult and the method is valid only for $V_p > 200$ volts. De Lussanet de la Sablonière[25] developed a graphical analysis of the tube characteristics which by successive approximations gives the desired result. However, severe restrictions are placed on the voltage range which can thus be used. Treloar[109] constructed a second tube, identical with the triode on which measurements were to be made, except for a suppressor grid between the grid and plate. He obtained s from the second tube in which the secondary electrons were suppressed.

A variation of the triode method was developed by Treloar[55] in which the yield from a target in filament form can be measured. The filament was mounted near the cathode inside the grid of a cylindrical triode. By a graphical analysis of the tube characteristics, he obtained values of the yield from the filament target which checked quite well with yields obtained by other methods on flat targets. For large V_p, δ so determined will be too large since not all the primaries strike normal to the surface.

The majority of the yield measurements are now done by the electron gun method. Here a well defined narrow electron beam is formed and passes through a hole in the collector so as to strike the target. The secondaries so produced are taken up by the collector which surrounds the target. By careful collimation it can be arranged that practically no primaries strike the collector. As discussed in §1.17, some difficulty may be experienced due to primaries reflected with or without energy loss from the target. This method allows a nonuniform target to be explored by deflecting the primary beam. In general, the results are more easily interpreted than in the triode method. It requires a more complicated tube, and in particular, careful electron gun design, to give a well collimated, monochromatic primary beam with a constant focus spot over a wide range of primary voltage.

For measurements of the effect on the yield of various parameters other than V_p, an especially accurate technique is available. In an electron gun tube the target current is measured and V_p adjusted until the target current is zero; i.e., $\delta = 1$. The parameter, such as temperature, is then changed and V_p adjusted to give zero target current again. The actual change in δ must be obtained from an independent determination of the slope of the yield curve around $\delta = 1$. This has the advantage common to all null methods that very high amplification can be used in the measurement circuit itself.

Three methods have been used for measurements of the energy distribution of the secondaries; retarding electric field, transverse magnetic field, or longitudinal magnetic field which are abbreviated respectively to RE, TM, LM. In the RE method, the collector in an electron gun tube is run negative with respect to the target and the secondary current is plotted as a function of the collector to target potential. The first differential of this curve is the desired energy distribution. It should be emphasized that the collector should be spherical in shape, with the target of relatively small size, at the center. Otherwise a true energy distribution will not be obtained; e.g., if target and collector are parallel planes, only the normal component of velocity is measured.

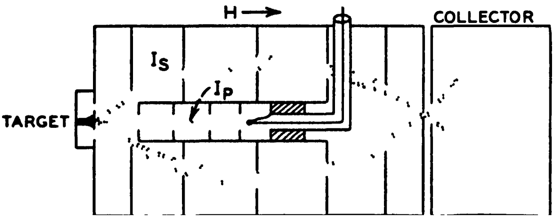


FIG. 9.—Measurement of secondary electron velocity distribution with a longitudinal magnetic field.[163]

In the TM method, electrons coming off the target in some given direction are fed into a magnetic analyzer consisting of a semicircular apertured tube in a uniform transverse magnetic field. The current to a Faraday Cage collector at the output end of the analyzer is obtained as a function of the magnetic field and this determines the energy distribution directly. In general this is more accurate than the RE method as it does not depend on a differentiation of the results. However, the TM method measures only those electrons which are emitted in one specified direction while the RE method includes all the emitted secondaries. For a complete description of the TM method, refer to Rudberg.[50]

Recently Kollath[163] employed another method originally used by Klemperer for β-rays. Termed the LM method, a magnetic field is parallel to the primary beam. Apertures are set up so that all secondaries emitted within a certain cone enter the linear analyzer. The axial focusing properties of the magnetic field are used to select secondaries of a given velocity which are measured in a Faraday Cage. As in the TM method, the relation of the Faraday Cage current to the magnetic field gives the energy distribution directly. The LM method allows many more secondaries to be used in the measurement than in the TM method.

Hence a smaller primary current may be used giving a greater protection against space charge effects. However, it does assume that the energy distribution is the same in all directions making a given angle with the normal to the target surface. As far as is known this assumption is correct for polycrystalline targets although it might not be valid for some coarsely crystalline targets.

All energy distribution measurements require the elimination of stray fields and of space charge effects and corrections for contact potential. Thus the experimental techniques are much more difficult than those involved in simple yield measurements.

II. Insulators

2.1 Secondary Emission from Insulators

Insulators are normally defined as those substances which have resistivities in excess of about 10^6 ohms-cm. at room temperature. Any insulator, which can be raised to a sufficiently high temperature without changing form, can be classed as a semiconductor. If the yield depends markedly on the conductivity, this presents a certain ambiguity which is, however, no worse than some others which occur in this phase of SE. The number of variables involved in the SE from insulators is much greater than for metals and so are the experimental difficulties. Consequently, apparent contradictions in results and their interpretations by various authors are not uncommon. Since much of the available information cannot be simply classified, this section cannot contain descriptions of all recent work but rather presents those results which, it is hoped, will be of most general interest.

2.2 Sticking Potentials and Yield

Most insulators have a δ/V_p curve similar in shape, although not necessarily in magnitude, to that for metals in which, for a certain range of V_p, the yield is greater than unity. There will thus be two bombarding voltages for which the yield is unity. These will be designated at V_p^{I} and V_p^{II} corresponding to positive and negative slopes respectively of the δ/V_p curve. These points assume greater importance with insulators than they do with metals because if an insulator is bombarded by a continuous flow of electrons, these points serve to separate three different types of operation. If V_p is less than V_p^{I}, the surface charges negatively until it approaches the cathode potential when no more primary electrons can strike it. If V_p is greater than V_p^{I} but less than V_p^{II}, the surface will charge positively up to a potential nearly equal to that of the collector such that the space charge reduces the effective yield to unity. If

TABLE III. SE yield of various insulators.

Insulators	δ_{max}	V_p max volts	V_p^{I} volts	V_p^{II} k volts	References
Glasses					
Pyrex	2.3	400	....	2.4	244
Pyrex	2.3	340	< 40	2.3	170
Nonex	....	....	....	3–5	136, 103
Soda	2.1	300	....	0.90	159
Cover	1.9	330	< 60	1.7	170
Ground	3.1	420	....	3.8	170
Quartz	2.1	400	30	2.3	170
Quartz	2.9	440	< 50	2.3	170
Phosphors					
Willemite	....	....	....	3–7	80
Willemite	...	....	. . .	5–10	136
Willemite	.	..	...	20	103
Zinc sulfides	.	.	. .	6–9	136
Calcium tungstate	.	..	.	3–5	136
Alkali Halides					
LiF	5.6	....	.	....	120
NaF	5.7	....	.. .	....	120
NaCl	6.8	....	....	....	120
NaCl	....	...	20	1.4	125
NaCl	6	600	....	....	147
KCl	7.5	....	....	....	120
RbCl	5.8	....	....	....	120
CsCl	6.5	....	....	. ..	120
NaBr	6.2	...	....	...	120
NaI	5.5	....	....	....	120
KI	5.5	....	...	....	120
Alkaline Earth Compounds					
CaF_2	3.2	..	....	....	120
BaF_2	4.5	. .	.	.. .	120
BeO	3.4	2000	.	.	218
MgO	2.4	1500	..	. .	218
MgO	4.0	400		.	87
CaO	2.2	500	..	.	218
SrO	2.6	500	.	. .	218
BaO	2.3	1600	.	. .	218
BaO	4.8	400	.	.. .	87
Oxide cathode (BaO, SrO)	8	1500	60	3.5	252
Oxide cathode (BaO, SrO)	5–12	1400	40	....	257
Miscellaneous					
Al_2O_3	1.5	400	...	1.7	159
Al_2O_3	4.8	1300	.	....	218
Al_2O_3	2.5	350	.	..	87
Al_2O_3	....	....	20	1.2	125
Mica	2.4	300	....	1.7	159
Mica	2.4	380	30	3.3	170
Mica	....	....	....	3.5	172
Mica	....	..	20	1.0	125

V_p is greater than V_p^{II}, the surface will charge negatively until the yield is unity; i.e., the surface potential becomes equal to V_p^{II}. In many actual insulators, V_p^{II} is to some extent a function of the collector voltage; i.e., as the collector voltage is increased in excess of V_p^{II}, the latter also increases slowly. Nelson[103] concluded that in the absence of a strong field at the surface, negative charges on various parts of the surface exert a "grid" effect which prevents other secondaries from escaping. This will be discussed more fully in §2.3.

Since the unity yield voltages or sticking potentials form the limits for different types of behavior which are important in practice, it is desirable when specifying the yield of insulators to include values for V_p^{I} and V_p^{II}. Table III gives a summary of most of the available information on the maximum yield and the sticking potentials of insulators. These are all supposedly thick targets and as such the results should not include any thin film phenomena, although it is not always easy to ensure this latter condition. These results are also for room temperature or, at least, a temperature such that the yield is not expected to be appreciably different from that at room temperature. Where the "heat conduction" method is used (see §2.9 on measurement methods) this may not be a valid conclusion.

2.3 Saturation of SE Yield

In metals, it is only necessary to apply a collecting field at the surface of the target sufficient to overcome space charge effects in order to collect all the emitted secondaries. A possible exception to this is the case of a very rough target surface. In some insulators, however, it is frequently found[103,244,119] that a considerable field strength is required to obtain saturation of the secondary electron current. Nelson[103] suggested that owing to inhomogeneities in the distribution of potential on the surface, some areas are effectively shielded by others which thus produce a "grid" effect, and that to penetrate to these shielded regions a higher field is required. Such an effect is also created by the finite conductivity of the target if this has not been taken into account by the method of measurement, as discussed in §2.4. Bruining[119] has suggested an alternative mechanism for certain cases based on his observations that a certain time is required for the secondary current to build up to its full value. He proposed that small surface particles become highly charged under the influence of the field and thus secondaries are drawn through the particles with sufficient velocity to create tertiary electrons. This is essentially one hypothesis of "field enhanced emission" which will be discussed more fully in §2.7. Where a definite saturation is finally obtained, it appears probable that the former explanation holds although it is conceivable that both effects could take place simultaneously. If field enhanced emission is involved it is possible that at very high field strengths it would change to thin film field emission (Malter effect).

2.4 Effect of Temperature and Conductivity on SE Yield

Unlike the situation with metals, the yield from insulators may change considerably with temperature if processes are involved which depend on the conductivity. However, if these are not present, if chemical changes are not involved, and if the methods of measurement are not sensitive to changes in conductivity, there is evidence that the yield from at least some insulators is temperature independent. Mueller[244] found that the unsaturated or "apparent" yield from pyrex glass was temperature dependent but that the true yield was independent of temperature over the range studied of 100 to 385°C. Vudynski's[147] work on NaCl and KCl although rather indefinite, indicates no appreciable change in yield on heating to 250°C.

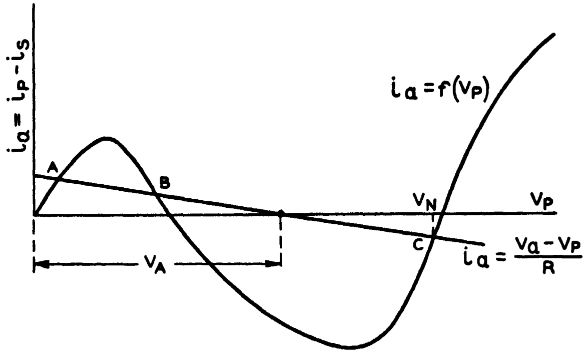


FIG. 10.—Effect of target conductivity on surface potential.

It is evident that both the "grid" effect and field enhanced emission discussed in §2.3 should disappear as the temperature, and with it the conductivity, is increased. However, in the former case the yield should increase with temperature and in the latter case it should decrease. If these were the only temperature dependent processes involved, they could thus be distinguished. Another important effect of the conductivity variation occurs if conventional methods of yield measurement are employed which neglect the effect of the finite conductivity of the target. This is illustrated in Fig. 10 where, following Bruining, we have drawn the dynatron characteristic of the target surface; i.e., the current through the target which equals the difference between the primary and secondary currents, plotted against the energy of the primary electrons V_p as they strike the target. The exact shape of the dynatron characteristic as $V_p \to 0$ will be determined by the reflection coefficient and by space charge effects in the primary beam. In any case $i_a = 0$ for $V_p < 0$ so that the exact shape at the origin will not affect the following argument. If the target has a finite resistance R, we can also represent the current through it by Ohm's law; i.e., $i_a = \frac{V_a - V_p}{R}$ where V_a is the potential between the cathode and the electrode on the target. Such a line may cut the dynatron characteristic at three points represented by A, B, and C of which B is unstable. If the collector voltage is sufficiently high, there are then two points of stable equilibrium; one for which $\delta \sim 1$ and the other where $\delta \sim 0$. The initial potential of the surface determines at which point the surface will remain. Of course, the target surface cannot exceed the potential of the collector. If the collector potential is less than V_N and i_a is negative, i_s/i_p will be greatly dependent on the collector potential since the target surface potential will tend to follow the collector. In a true insulator where $R = \infty$ it is obvious that special methods of measurement such as discussed in §2.9 must be used. However, this is not always obvious with targets of finite resistivity where standard methods are frequently used. In these cases the effects of target resistivity should be carefully analyzed.

The oxide coated cathode. Owing to its importance technically, the variation of yield with temperature of the oxide coated cathode has been studied in considerable detail. Such a target consists of a thin layer (about .001-inch thick) of a mixture of barium oxide and strontium oxide. At room temperature it is a fair insulator with a resistivity of about 10^8 ohm-cm. This decreases to about 10^3 ohm-cm. at 800°C., the normal operating point as a thermionic emitter. However, the resistivity is a function of the degree of activation and this may introduce some degree of uncertainty as to the validity of close comparison between the results of different workers.

Morgulis and Nagorsky[101] used a standard electron gun measuring technique and relied on the use of small currents to eliminate effects due to the layer resistance. Their highest value of V_p was 1200 volts which was insufficient to reach $\delta_{\max}$. At $V_p = 1000$ volts they obtained $\delta = 3$ which is much lower than that obtained by Pomerantz or Johnson at that voltage for a well activated target. As the temperature of the target was increased they found that the yield increased exponentially, attaining a value of $\delta = 12$ for $V_p = 1000$ volts at 850°K. where the thermionic emission was considerably larger than the secondary emission. They fitted their curves to the expression

$$\Delta\delta = Ae^{-Q/(2kT)} \tag{17}$$

where $\Delta\delta$ = increase in yield over that at room temperature
$Q = 0.70$ electron volts and is independent of V_p.
They also measured the energy distribution of the normal component of velocity of the secondaries and concluded that the most probable energy, which equalled 4 volts at room temperature, decreased to 2.3 volts at 600°K. Finally, they concluded that further work must be done to ensure that the results had not been affected by the target resistance.

Pomerantz,[252,253] using a DC triode method, obtained values for $\delta_{\max}$ ranging from 3.2 to 6.8 at room temperature for $V_p \doteqdot 1500$ volts and indicated that targets of low activation have substantially lower yields than those with normal activation. He observed an exponential increase in the yield as the target temperature was increased and fitted the results to the same expression as that used by Morgulis and Nagorsky. His values for Q ranged from 0.8 to 1.5 ev. Although thermionic emission was too copious at 850°C. to enable SE measurements to be made, he assumed that the exponential increase in δ observed at lower temperatures would be valid at 850°C. and thus obtained extrapolated values of δ at that temperature ranging from 84 to 136. Such DC measurements do not permit the determination of short time yield variations. To overcome this objection, Pomerantz conducted some pulsed measurements as described in §2.9. Operating in the millisecond range, he could not observe any variations in yield with time which might indicate Malter effect. Using microsecond pulses, he observed nothing abnormal until the target reached such a temperature that it was emitting an appreciable thermionic current. Then he observed the same behavior as that ascribed by Johnson to "Bombardment Enhanced Thermionic Emission" which is described later. Pomerantz obtained similar behavior on bombarding a tantalum target that had been heated by a separate filament so as to emit thermionically. From this he concluded that the effect observed by Johnson "depends critically upon certain experimental factors including the geometrical disposition of components of the experimental tube and is a property to be associated with space charge rather than with the target itself." Pomerantz obtained good agreement of yields measured by the DC method with those observed with the pulsed method and stated that significant increases in δ occurred before the onset of detectable thermionic emission from the target.

Pomerantz also made some retarding potential measurements which, from his tube geometry, must be interpreted not as true energy distributions of the secondaries but rather as some function thereof. He observed a decrease in the peak value of this function as the temperature increased, which he interpreted as being in agreement qualitatively with the results reported by Morgulis and Nagorsky.

We have considered Pomerantz's results in some detail because they contain some substantial points of disagreement with those of Johnson, supposedly on the same type of target. All of Johnson's reported work[235,248,249,257] has been done with a pulsed technique using pulses of a few microseconds duration. At room temperature, he obtains values of $\delta_{\max}$ for $V_p \div 1400$ volts ranging from 5 to 12, depending on the condition of the target. Defining the yield as that obtained from the peak value of SE current during the pulse, he obtains a variation in yield with temperature as shown in Fig. 11. The waveforms shown are those of the current through the target corresponding to the different temperature ranges. At low temperatures, the current charges up the target surface, causing it to approach the collector potential and thus reducing the effective yield, near the end of the pulse. This effect was not observed by Pomerantz, which suggests that his targets had considerably higher conductivity. As the temperature increases, the yield decreases, levelling off around 500°C. In this temperature region, the conductivity has increased sufficiently to eliminate charging up of the surface. When

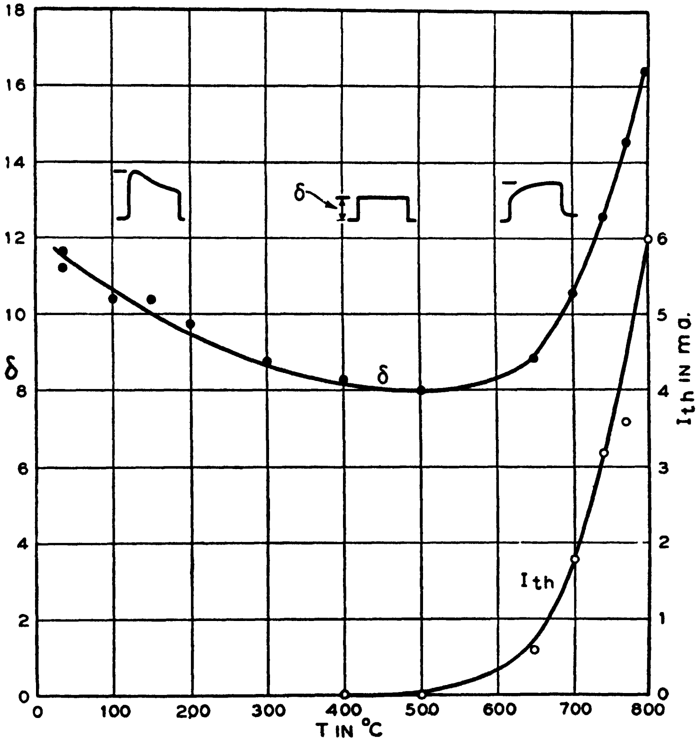


FIG. 11.—Variation of yield of oxide coated cathode with temperature.[257]

thermionic emission sets in, the yield increases sharply but the pulse shape shows that this increase is due to electrons which are not emitted simultaneously with the arrival of the primaries but increase in number with time more or less as $1 - e^{-t/\tau}$ and continue to be emitted after the primary beam is turned off, finally dying away after a few microseconds. If $\Delta\delta$ is defined as the difference between the yield at high temperatures and that at 500°C., Johnson can fit his results to the same equation used by Morgulis and Nagorsky. He interprets the observed yield at high temperatures as being due to a true secondary yield with a δ probably equal to that obtained at 500°C., and, superimposed on this, a temporary enhancement of the thermionic emission caused by a change in the thermionic emission constants as a result of the electron bombardment.

Johnson has shown that the enhanced part of the yield appears to vary linearly with primary beam current (as does the true SE yield), that it is practically independent of bombarding voltage above $V_p = 200$ volts which is radically different from the behavior of the true yield, and that the enhanced yield is roughly proportional to the steady thermionic current. He has obtained these results with a number of targets, with different tubes of differing construction geometrically, and over a range of primary currents. It is difficult to imagine any space charge mechanism, such as was suggested by Pomerantz, which could account for such behavior under these widely differing circumstances, and it appears probable that Johnson's interpretation is essentially correct. If we do not accept it, we are faced with the formidable problem of formulating some other mechanism which can account for the very high yields observed at high temperatures. At present, there is no theoretical justification for such yields from a target with high conductivity. However, when we accept Johnson's interpretation, we have the difficulty of accounting for the increase in yield with temperature at lower temperatures observed by Morgulis and Nagorsky and by Pomerantz, as contrasted with the decrease observed by Johnson. We can only surmise that the target materials were not the same. It is tempting to suppose that Johnson's high yields at room temperature were caused by field enhanced emission which would decrease as the conductivity increased. This will be discussed further in §2.7.

To sum up, it is evident that the effects of temperature on the yield of insulators may be quite complicated and it is probable that only through the use of pulsed techniques can the various processes involved be evaluated properly.

2.5 Velocity Distribution of Secondary Electrons from Insulators

Geyer[217] has measured the velocity distributions of slow or "true" secondaries (see §1.15) for layers of various thickness of NaCl and MgF_2

on a base of nickel by a DC retarding field method. For films sufficiently thin to give adequate conduction, he found that the most probable energy of the secondaries $V_{s\ max}$ was equal to 1 volt or less.

Vudinsky[149] has also investigated the velocity distribution of sodium chloride by a similar method. He showed that the distribution is the same for a single crystal as for a powdered layer and that $V_{s\ max}$ is essentially independent of the bombarding voltage V_p although the average energy of the secondaries does increase slightly as V_p increases from 300 to 1500 volts. He obtained values of $V_{s\ max}$ which were just below 1 volt although no correction was made for "contact" potential. To obtain adequate conductivity, Vudinsky heated up his targets. However, both he and Geyer observed a substantial increase in the yield as the potential of the collecting sphere was made positive with respect to the target; i.e., lack of saturation. Vudinsky observed that although a change in primary current altered the shape of the distribution curve it did not affect the position of $V_{s\ max}$. Thus, although it is probable that both experiments were affected by the lack of conductivity, yet the values of $V_{s\ max}$ may be reasonably reliable.

Johnson[257] has shown that the mean velocity of emission of secondaries from an oxide cathode is much less than that from nickel. We may conclude that at least for those few insulators which have been studied, the mean energy and the most probable energy of emission of secondaries from insulators are considerably less than from metals. This is in agreement with conclusions reached by Kadyschevitsch[242] based on an extension of his theory of SE from metals and dielectrics.

Very little is known about reflected primaries from insulators. Bruining[89] has shown that in the range of $3\ \text{volts} < V_p < 25\ \text{volts}$, the yield or reflection coefficient from barium oxide is several times that from barium. Krenzien[220] has obtained similar results on thin layers of alkali halides. He also concluded that the reflection was completely elastic until V_p exceeded a value which corresponded to the long wavelength limit of ultraviolet absorption of the crystal. Rudberg[50] observed loss peaks of BaO and CaO which differed from those of barium and calcium but no data are available concerning the yield of reflected primaries, including all, those which have been inelastically reflected for medium or large values of V_p.

2.6 Miscellaneous Properties of SE from Insulators

Practically no information is available on the range of slow electrons in insulators. On indirect evidence, Geyer[217] concluded that electrons with less than 1000 ev energy could travel more than 300 molecular diameters in NaCl. Bethe[189] and others have suggested that the mean free paths of secondaries in insulators should be much greater than in metals since there are relatively few conduction electrons to which they can lose energy. Thus to be absorbed they must recombine with a positive hole or be trapped by an impurity or imperfection. The fact that the maximum yield from insulators usually occurs at a higher V_p than for metals suggests that this may be correct if the depth of penetration of the primaries is roughly the same in both cases.

In some cases the apparent yield from insulators is not independent of primary current. Where surface charging effects take place, the time rate of change of surface potential will be a function of the primary current. Moreover, Bruining[80] has shown that the yield from alkali halides, in particular sodium chloride, decays to lower values with time of bombardment. Presumably this is due to the creation of color centers by the primary bombardment. Eventually metallic agglomerates are formed which effectively change the target composition. The rate at which they are formed is again a function of primary current. Vudinsky[212] has reported an increase in yield of KCl and NaCl with increased primary current. It is probable that this was due to the additional conductivity introduced by the color centers which increased the apparent yield as distinguished from the true yield. He has also "stabilized" the yield from these insulators by exposure to vapors of alkali metals, thus creating more color centers.

It was once thought that the variation of yield with angle of incidence for insulators was quite different from that for metals, involving certain critical angles where abrupt changes in yield occurred. This has since been attributed to surface charging phenomena. Recently Salow[171] and Scherer,[172] using quite different methods, have shown that for mica, glass, and ZnS the variation of yield with angle of incidence agrees well with Müller's data on metals.

2.7 Double Layer Formation and Field Enhanced Emission

The hypothesis of field enhanced emission has been proposed to explain the results obtained in a number of experiments with insulators and insulating films. Essentially, it is assumed that a field is created by positive surface charges or by other means which extends into the body of the insulator and thereby increases the yield either by creating an "avalanche" of electrons or by lowering the surface barrier in the vicinity of the surface charge. The effect is likened to an "inertialess" Malter effect (see §2.4); i.e., the secondary current does not exhibit the time lags characteristic of Malter effect. The hypothesis of field enhanced emission has been used to explain the very high yields obtained from some insulators, lack of saturation of secondary current with collecting field and certain discontinuities in the yield of insulators around unity yield points. Inasmuch as alternate hypotheses can be advanced to explain at least some of these effects, it is a somewhat controversial subject and all the evidence for or against it cannot be presented here. Instead, some representative experiments and arguments will be discussed. The experimental evidence is considered in greater detail in a recent article by Trey.[281]

Hintenberger[125] bombarded targets of mica, Al_2O_3 and NaCl using a DC technique and measured the surface potential as a function of time. For V_p slightly greater than $V_p{}^1$ so that $\delta > 1$, he observed that at first the surface charged up close to the collector potential. It then fell off slightly and finally fell abruptly to the cathode potential. For higher values of V_p, the surface remained near the collector potential throughout. He proposed that while the surface charged positively since $\delta > 1$, the primary electrons themselves formed a negative space charge underneath the surface. Thus a field was formed tending to expel electrons from the target, thereby enhancing the yield. Continued bombardment resulted in a movement of the internal space charge towards the surface since further primaries would not penetrate as deeply due to space charge repulsion. Finally when the space charge reached the surface, it would fall to cathode potential. It should be noted that this type of behavior could also be explained by a "patch" field theory in which part of the surface initially had a $\delta > 1$ and part with $\delta < 1$. As the low potential area charged towards cathode potential it would bombard the high potential area with slow electrons for which $\delta < 1$ and eventually the entire surface might fall to the cathode potential. This is probably not a very good hypothesis but at least it is a possible one which does not involve field enhanced emission.

Nelson[104,167] observed a sudden increase in yield with a film of MgO on nichrome as V_p increased through $V_p{}^I$ where $\delta = 1$. He attributed this to field enhanced emission and showed that it coincided with an abrupt increase in the surface potential. Yasnopol'ski[186] has attributed this action to the layer resistance itself (see §2.4) although Nelson believed that the relation between the net flow of current through the film and the voltage across the film was nonohmic. Copeland[122] has also observed a sudden break as V_p increased through $V_p{}^{II}$ for gassy sodium films which he attributes to double layer formation. Unfortunately he did not investigate the variation of surface potential simultaneously.

The importance of the hypothesis of field enhanced emission becomes evident when an explanation is sought for the high yields found in many insulators. Bruining and de Boer,[120] Maurer,[202] and Morgulis[164,165] believe that such high yields occur in insulators in which the first empty conduction band lies near or above the potential of the surface barrier (see §2.8) and that there is no special emission from so-called active centers, or enhancement of the yield due to fields formed in the insulator.* On the other hand, Timofeev,[175–181] Pyatnitski,[175] Frimer,[156] and others believe that, in all such cases, positive ions imbedded near the surface help to draw the electrons out in a modified form of Malter effect. Even though the lifetime of such ions before recombination may be very short, they may still help many electrons to escape. Timofeev and his coworkers have carried out a long series of experiments with complex surfaces, the results of which they have explained by this hypothesis. However, they have not used pulse techniques and their results are complicated by the effects of target conductivity so that alternative explanations of their results are applicable. In any case their ideas, right or wrong, have enabled them to prepare high yield surfaces of considerable stability which will be discussed in §3.3.

One of the main difficulties with these various hypotheses of field enhanced emission is that there is not enough information available to put them on a quantitative basis. The validity of Timofeev's assumption that a positive surface charge augments the yield by lowering the work function, cannot be evaluated until more is known about the effect of work function on the yield from insulators. Although the work function plays a relatively minor role in determining the yield from metals, this may not be the case with insulators, especially those in which the bottom of the conduction band lies very close to the top of the surface barrier.

Another unknown factor is the magnitude of the field strength required across all or part of the target to produce an appreciable increase in yield. If this field serves merely to produce a drift velocity of secondaries toward the surface, it may not have to be very large since the secondaries may have a long life time in the conduction band. If, however, the field strength must be sufficient to produce tertiary electrons, as assumed by Bruining, it must be close to that required to induce dielectric breakdown, namely, $\sim 10^6$ volts/cm. In any of the proposed mechanisms of field formation, a finite time is required to set up the field and this suggests a possible approach to the problem. Let us examine the variation with time of the field strength E set up across a target by the production of a net positive charge on the bombarded surface with $\delta > 1$. If the primary electron penetration is negligible compared with the target thickness, then the field strength at any given instant of time "t" seconds after primary bombardment begins, is given by

---

* It should be noted that Bruining does admit the possibility of the existence of field enhanced emission in those cases where the secondaries exhibit a pronounced lack of saturation with collector voltage. He suggests that the secondaries may acquire sufficient energy from the field inside the insulator to produce tertiary electrons.[119]

$$E = (\delta - 1)\rho j_p(1 - e^{\frac{-1.13\times 10^{13}t}{\rho K}}) \text{ volts/cm.} \tag{18}$$

where ρ = specific resistivity in ohm-cm.
j_p = primary current density in amp./cm.^{2}
K = dielectric constant.

This field strength will continue to increase with time until the surface potential approaches that of the collector and space charge effects set in, or until the leakage current density through the target equals $(\delta - 1)j_p$ if space charge is negligible. Thus E cannot exceed either V_c/d or $(\delta - 1)j_p\rho$ where V_c is the potential of the collector relative to the target electrode and d is the target thickness. If the leakage current is neglected, eq. 18 reduces to

$$E = \frac{(\delta - 1)}{K} j_p t \cdot 1.13 \times 10^{13} \text{ volts/cm.} \tag{19}$$

If, instead of assuming that the field is effective throughout the target thickness, we accept Hintenberger's hypothesis that the field is formed between the positive surface and the trapped primary electrons, we would modify eq. 19 by replacing $(\delta - 1)$ by simply δ.

The possibility was mentioned earlier that the high yields which Johnson observed for oxide cathodes at room temperature might be caused by field enhanced emission. In a typical measurement he has observed that the SE pulse reaches its maximum value in less than 10^{-7} seconds. The other required values are $\delta = 11$, $j_p = 1.5 \times 10^{-4}$ amp./cm.^{2}, $K = 3.5$. Putting these in eq. 19, we find that the field strength across the thickness of the target at maximum yield is 480 volts/cm. If we assume that trapped primaries augment the field, we could increase this to 530 volts/cm. These would appear to be rather low field strengths compared to those required in photoconductivity work. It is unlikely that such low fields could affect the yield appreciably unless the bottom of the conduction band lies very close to, or above, the surface barrier since the secondaries will rapidly lose energy until they reach the bottom of the conduction band. The field will then produce merely a drift velocity in the direction of the surface and if the surface barrier cannot be traversed by electrons with thermal energies, they cannot contribute to the SE yield.

2.8 Theories of Secondary Emission from Insulators

Apart from the hypothesis of field enhanced emission which has already been discussed, several theories, mainly qualitative, have been proposed. Bruining and de Boer[120] assume that if the first conduction band available to the secondary electron is near or above the potential of the surface barrier, the secondary has an excellent chance to escape into vacuum, particularly since it probably has a very long mean free path. However, in a semiconductor possessing allowed energy levels which are well below the surface potential, a secondary electron may go to the lower level in which case it cannot escape. It may also go to a higher level, but then it has the possibility of dropping down to the lower unfilled levels thus reducing the probability of escape. They give the alkali halides as examples of the first group ($\delta \sim 5$), and compounds such as Cu_2O, MoS_2, ($\delta \sim 1$) as representative of the second. The latter usually show light absorption at longer wavelengths than the "red" limit of the external photoelectric effect. Bruining and de Boer also showed that for NaCl the actual emission was ten times the rate of formation of color centers by the primary beam. From this they concluded that the emission does not originate from active centers but from the lattice electrons themselves.

Maurer[202] has fitted these ideas into a more quantitative form but since his final equation for the yield involves a number of quantities, such as diffusion coefficients, which are at present unknown, it is difficult to make use of it except qualitatively and, in that sense, he is in agreement with Bruining and de Boer.

Kadyschevitsch[196] has extended his theory of SE from metals to include dielectrics and semiconductors. The treatment is extremely similar except for different parameters. In particular, he assumes that the secondaries must receive a large impulse from the primaries and hence the interaction can be considered as between two free electrons. The final conclusions are again essentially qualitative, since they depend on unknown parameters, and are in general agreement with Maurer.

As is the case for metals, little attention has been paid to the precise role of the surface in any of the theories of SE from insulators. The possible effects of surface states[86] has not yet been treated. There is no evidence that the surface is less important in insulators than in metals and it is indeed possible that here the surface effects are of great importance.

2.9 Methods of Measurement of SE for Insulating Targets

Measurements of the yield from insulators require either that the potential of the bombarded surface be fixed or else that it be measured simultaneously with the SE measurement. The means of accomplishing one or the other will be subdivided into static or quasi-static methods and dynamic methods.

Static Methods

Low current method. Most of the earlier work was done using low values of DC current in either the triode or electron gun method. Although the bombarded surface charges up (or down, depending on whether δ is greater or less than unity), the surface potential changes slowly and the yield can be plotted as a function of time. By extrapolation the actual yield can be deduced from this apparent yield. The inherent limitation to extremely small currents can be relieved if the surface potential can be controlled by some other means such as those that follow.

Heat conduction method. Upon heating an insulator, it becomes either an intrinsic or impurity semiconductor providing thermal decomposition does not occur. Metallic contact can be made to the outside target surface; i.e., the surface opposite to that which is bombarded. If the input resistance of the measuring circuit is much higher than the resistance through the semiconductor, the voltage drop through the target will be negligible. Nelson[103] and others[123,136] have used this method in the study of SE from luminescent screens inside cathode ray tubes. If a high sensitivity electrometer is used for the measurement, the glass envelope need not be heated excessively. The method was also used by Mueller[244] in a study of pyrex glass.

Thin film method. Geyer[217] and others have used very thin insulating films on a metallic base such that electrons could be drawn through the film and thus control the surface potential. However, it is doubtful if the properties of such thin films are representative of those of the bulk material.

Excitation by infrared. In studies on sodium chloride, Geyer[217] produced color centers by the electron bombardment. He then irradiated the target by infrared light, ejecting electrons from the color centers into the conduction band. These electrons then acted to neutralize the surface charge. However, the results are difficult to interpret since the density of color centers as a function of primary electron path length must be known.

In any conduction method it is evident that the concentration of electrons in the conduction band should remain low or else the yield may not be characteristic of that of the original insulator.

Direct measurement of surface potential. There is some doubt if an electrostatic measurement of the surface potential of an insulator is always valid because of the possibility of the formation of a charge layer under the surface.[125] However, the evidence for such behavior is not as yet conclusive. The following methods of surface potential measurement have been used.

Electron beam voltmeter. Scherer[172] and Piore and Morton[168] bombarded the target with an electron beam directed normally to it. A second electron beam was shot at right angles to this beam, parallel to the target surface and a few millimeters above it. The second beam was deflected by the surface potential of the target, the deflection being observed on a fluorescent screen. A deflection calibration was obtained by replacing the insulator by a metallic target and plotting the deflection as a function of target potential.

Grid method. Nelson[167] mounted an auxiliary filament and a collector close to the target surface and used the target surface potential as a "grid" to control the electron flow from the filament to the collector.

Velocity distribution method. Nelson[103] also mounted a Faraday Cage collector in such a way as to intercept secondaries from the insulator but presumably such that voltage variations of the cage would not affect the target surface potential. Plotting the cage current as a function of cage voltage, he obtained a sharp "break" where the cage voltage equalled that of the target.

Direct measurement method. Nottingham[80] embedded some fernico wire probes in the insulator near the bombarded surface and thus measured the surface potential.

Dynamic Methods

One of the reasons for the relatively small amount of work which has been done on SE from insulators is the difficulty or uncertainty of the static methods just described. It appears probable that the application of high frequency or pulse techniques, which are classified here as dynamic methods, will do much to develop this important field.

Direct pulsing method. Although the most recent historically, this is essentially the most basic of the dynamic methods. It was first used by Johnson[235,257] and later by Pomerantz[252] in the study of oxide coated cathodes. Fig. 12 shows Johnson's circuit schematically. The primary electron source is maintained at $-V_p$ but the beam is normally cut off by the negative grid bias. The back of the target is connected through a small resistor to ground and the voltage developed across it is amplified by a wide band video amplifier and displayed on an oscilloscope. Due to the small but finite conductivity of the target, the target surface is normally at ground potential. The pulser then delivers a short ($\sim$ 1μsec.) flat topped pulse to the grid which turns on the primary beam for this interval. This is synchronized with a fast horizontal sweep on the oscilloscope. The impedance of the target layer itself thus produces a current through the load resistor equal to $i_s - i_p$ which appears as a pulse on the oscilloscope. If the primary current and the target conductivity are such as to cause appreciable surface charging during the duration of the pulse, this will be observed as a reduction in pulse height towards the end of the pulse. By varying the recurrence frequency of the pulses i.e., the duty cycle, a sufficient time can be allowed between pulses to enable the target surface again to reach ground potential. The primary current can be measured by observing the pulse with the collector negative as shown or by observing the current pulse leaving the cathode if the beam is well focused on the target. The latter method eliminates the effect of high-speed secondaries as discussed in §1.17. The measurement of δ can be made quite accurately by the use of well designed attenuators in conjunction with the amplifier.

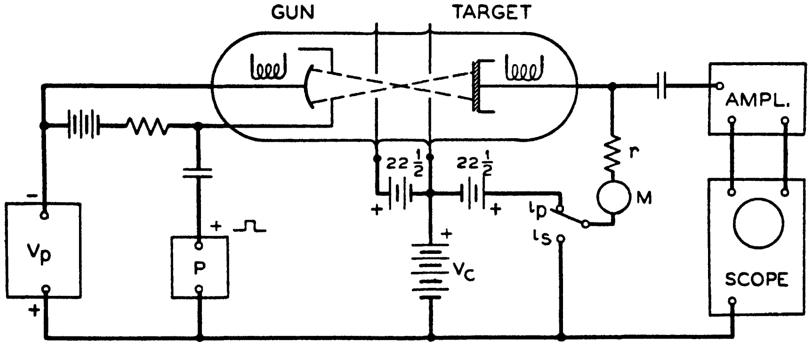


FIG. 12.—Schematic for SE yield determination by direct pulsing method.

Dynamic two gun method. Salow[170] used essentially the same method that Johnson did except that he used long pulses and the very high recurrence frequency of 50 kc. In this case, the surface could not recover to its rest potential before the occurrence of the next pulse so Salow flooded the target with an intense DC electron beam from a second gun at such a potential that $V_p(\mathrm{dc}) < V_p^{\mathrm{I}}$ and thus the target surface remained essentially at the potential of the cathode of the DC emitter. A somewhat similar method has been used by Nemilov.[207]

Heimann and Geyer's method. Heimann and Geyer[159] maintained a steady beam current and pulsed the potential of the primary emitting cathode. To set the target potential initially, they located the upper crossover point V_p^{II}. With $V_p = V_p^{\mathrm{II}}$ as the initial state, they suddenly reduced V_p to a smaller value by pulsing the cathode. The response was displayed in the same way as previously described and they could thus plot out the yield curve between the two crossover points. A limitation of this method is that it can only be used where $\delta > 1$ and it also requires a more powerful pulser than that needed for grid modulation. As they used it, it also required that the point V_p^{II} be located.*

III. COMPOSITE SURFACES

3.1 Composite Surfaces and Thin Film Phenomena

This section has been separated from that on insulators because it is not at all clear that the SE properties of thin insulating or semiconducting films on a metallic base are the same as they would be for the bulk material, even if pulsed methods were used which prevented surface charging effects. Moreover, the surfaces to be discussed are usually so complex that no attempt has as yet been made to explain their action except qualitatively. Consequently, both the results and their interpretations differ considerably and the overall picture can be presented only in very general terms. Unfortunately, these are the surfaces which are, at present, of the most commercial value since they may have high yields and also sufficient conductivity to eliminate charging effects. With few exceptions, all of the results to be quoted have been obtained by DC methods and in some cases surface charging probably has played a role.

3.2 Yield from Photocathodes of the Form [Ag]-Cs_2O, Ag-Cs

A typical photocathode is formed by oxidizing a silver backing plate so as to give a layer of silver oxide several hundred molecules thick. If cesium is then admitted, it is completely absorbed by the silver oxide. The system is then heated (to 250°C.) and a reaction takes place between the silver oxide and the cesium giving cesium oxide and free silver. Some excess cesium is absorbed on the surface of the layer and a sensitive photoelectric surface results.[29] It was used by Farnsworth[27] as a SE target in a SE multiplier and has since been the subject of many investigations. In general, δ_{max} ranges from about 5 to 10 at values of V_p between 500 and 1000 volts. Weiss[58] has shown that the yield is relatively independent of the thickness of the Cs_2O layer although for a certain thickness it does reach a maximum about 20% higher than normal. He has also studied the effects of different base metals and showed that the highest yield, about 11, was obtained for silver and the lowest yield, 2.3, for gold with ten other metals in between. Penning and Kruithof[34] showed that the photoelectric sensitivity could vary considerably but the SE yield would remain essentially constant. Zworykin, Morton, and Malter[61] have shown that the amount of free cesium present for the highest SE yield is less than that needed for maximum photosensitivity. These results have been confirmed by Timofeev and Pyatnitski[58] who obtained the highest yield for an oxide layer of about 200 molecular diameters on a silver base. They observed the SE and photoelectric yield simultaneously as the target was heated and the adsorbed cesium atoms were evaporated off. The maximum SE yield was obtained at a point where the free cesium concentration was much lower than that for optimum photosensitivity. Moreover, the SE yield was much less dependent on the free cesium concentration than was the photosensitivity. They obtained similar results with cathodes formed with rubidium or potassium in place of cesium.[82] From this they concluded that while the photoelectrons arise from the surface, the secondary electrons arise from so-called emission centers consisting of pieces of alkali metal distributed throughout the oxide, and the escape from the target of the secondaries is aided by field enhanced emission.[177]

---

* A similar "manual" pulsing procedure has been used to measure the SE of iconoscope targets. See Janes and Hickock, *Proc. Inst. Radio Engs.*, **27**, 539 (1939).

Dobroljubski[42–44,67] observed that the SE yield correlated with the integral photosensitivity in the ultraviolet but not in the visible or infrared. He concluded that the secondaries arise either from the oxide layer or the base material. Treloar[109] obtained yields similar to Zworykin's using a base material of either silver or copper. He also prepared a target which did not contain free silver in the cesium oxide and obtained practically the same yields as previously. From this he concluded that the secondaries do not arise from metallic emission centers in the oxide as claimed by Timofeev and Pyatnitski. Borzyak[117] also arrived at the same conclusion as did Treloar. It should be mentioned that Treloar found that he could increase the yield from such targets by about 20% by "aging" them. This consisted of drawing a 1-milliampere space current from the target for several hours.

Kwarzchawa[48] observed a decrease in SE yield from a surface whose photoelectric yield had decayed and concluded that the secondaries must arise from the adsorbed cesium layer. As previously discussed, it is most unlikely that a monomolecular adsorbed layer could give rise to an appreciable SE yield. However, it is quite possible that the photoelectric exhaustion caused a change in conductivity and could thus affect the yield.

Khlebnikov and Korshunova[94] demonstrated that it was necessary to have free silver or free cesium atoms in the oxide layer to obtain a high yield and concluded that the secondaries arise principally from these active centers. However, they admitted that the role played by these free particles might be merely to increase the layer conductivity.

From all this mass of data, much of it conflicting, we can arrive at the important conclusions that the photoelectric sensitivity to visible light has nothing to do with the SE yield properties and that the secondaries arise from within the oxide layer. Bruining believes that the free metallic particles merely serve to improve the layer conductivity while Timofeev thinks they are electron sources and the emission is aided by a positive space charge. Experiments, which may show conclusively which point of view is more nearly correct, have not yet been performed.

3.3 Yield from Other Oxidized Targets

Certain alloys have been shown to give a very high SE yield following oxidation which is frequently hastened by subjecting the target to a high frequency discharge in an oxygen atmosphere. Zworykin, Ruedy, and Pike[215] have obtained yields as high as $\delta = 10$ with a silver magnesium alloy so treated. Such yields were somewhat unstable with time. However, a yield of $\delta = 7$ could be obtained which remained very stable after 200 hours aging. Yields of $\delta = 16$ have been obtained with this alloy by Friedheim and Weiss.[192] Malter[200] has made velocity distribution measurements on magnesium-silver alloy and reported values for the most probable velocity $V_{s\,\max}$ ranging from 5.5 volts for $V_p = 100$ volts to 4.0 volts for $V_p = 200$ volts. Such a dependence on primary voltage is most unusual. He employed a simplified magnetic analyzer and it is possible that the observed variation of $V_{s\,\max}$ might be caused by the method of measurement.

Gille[193] and Mathes[201] reported high yields for alloys of beryllium with nickel, cobalt, iron, copper, tungsten, and molybdenum, with the best results for beryllium-nickel. If the oxidation process is continued too long, instability and Malter effect result. Randenbusch[209] has observed that poisoning of such surfaces occurs if an oxide cathode is activated in the presence of the target. However, this poisoning can be removed by subsequent heat treatment.

Timofeev and Aranovich[176] describe surfaces of oxidized magnesium and oxidized barium. Optimum yields were obtained for layers about forty molecules in thickness. Such targets can stand much higher heat treatments than the photocathode type. With $\delta_m \sim 5$ at room temperature, it increases to $\delta_m \sim 7$ at 800°C. If heated much above this, it drops to $\delta_m \sim 3$ and remains there henceforth. These targets were prepared by evaporating the magnesium or barium onto the base metal in vacuum followed by subsequent oxidation. Aranovich[233] has also described surfaces of microdispersed MgO on nickel prepared by evaporating magnesium through an atmosphere of dry oxygen. He states that yields of $\delta_m \sim 80$ could be obtained which although showing time lag effects, responded sufficiently quickly to follow a 40 kc modulation of the primary beam. Yields as high as $\delta \sim 10{,}000$ were obtained but definitely exhibited Malter effect. Bruining and de Boer[119] had used this type of surface previously and reported yields of $\delta_m \sim 18$. Schnitger[173] believes that free magnesium is necessary in the oxide to obtain maximum yields. He has shown that MgO films are relatively insensitive to exposure to oxygen at room temperature although they are rapidly destroyed by water vapor.

Nearly all these high yield surfaces with yields greater than about 10 show a lack of saturation with collector voltage and a certain amount of instability which, if the oxidation is continued, may break over into Malter effect. The evidence suggests that the yields may be the result of field enhanced emission.

3.4 Malter Effect (Thin Film Field Emission)

In 1936 Malter[49] reported some remarkable results obtained with a cathode of composition [Al]-Al_2O_3-Cs_2O. These were formed by oxidizing an aluminum plate to give a film of Al_2O_3 about 2000 A. thick. He coated this with cesium, heated it, and oxidized it. With bombarding voltages of a few hundred volts he obtained emission currents from the target which were as much as one thousand times the primary current. He found that the emission current varied as a power of the collector voltage V_c and as a power of the primary current I_p, within certain limits of V_c and I_p. The emission current did not reach its full value until some time after the bombardment began. On shutting off the primary current, the emission current decayed rapidly at first, and then slowly approached the zero value asymptotically. This decay time could be very long; one surface exhibited a detectable emission 24 hours after cessation of the primary bombardment. He showed that the decay was greatly accelerated when the surface was irradiated by visible light. It was also found impossible to obtain completely reproducible results. At high values of V_c and I_p, scintillations were observed on the surface which altered the target characteristics; intense scintillations caused the phenomenon to disappear.

Malter interpreted these results as being caused by an action analogous to the "spray discharge" reported by Güntherschulze[23,24] in which he obtained a gas discharge without a negative dark space or cathode fall. Malter suggested that the surface has a yield greater than unity and that due to the high resistivity of the oxide layer, a positive charge is built up, upon bombardment, which eventually sets up sufficiently intense gradients to cause field emission from the aluminum and aluminum oxide. In addition, the oxide becomes polarized and both the polarization and the surface charge persist after the removal of the primary beam until neutralized by leakage and by a portion of the field emission. The relation between emission current and collector voltage is shown to be of a form similar to the current voltage characteristics for Thyrite. Güntherschulze has ascribed the nonohmic characteristics of Thyrite to field emission across oxide films which separate the particles of carborundum of which it is composed. The scintillations were presumably caused by actual breakdown at weak spots in the film. Malter also obtained scintillations from SiO, MgO and willemite and concluded that these should show thin film field emission. He obtained negative results with Ta_2O_5, CaO, Ag_2O, CuO, NiO_2, WO_2, and ZrO and concluded that, in the heat treatment, they were reduced by the cesium which greatly reduced their resistivity.

This phenomenon has subsequently attracted many investigators, partly because of its intrinsic interest and partly because it suggests a method of obtaining high yield surfaces of commercial value if it could be stabilized. Most of this work has confirmed Malter's observations and his interpretations, but has not yielded significantly improved surfaces from the point of view of stabilization and reproducibility.

Koller and Johnson,[73] and also Mahl,[77,100] have studied the effect by using the emitted electrons to form an electron-optical picture on a fluorescent screen. They showed that the surface does not emit uniformly but that the electrons come from distinct spots like pin points. New spots could be formed by making the emitted electrons return to the target by application of a magnetic field. They interpreted the scintillations as due to sudden bursts or eruptions of electrons which would tend to discharge the surface near the emission spot. Thus the surface potential changed continually and under the proper conditions it varied cyclicly. They showed that the "Malter" electrons are emitted with a wide range of velocities, and that some had an energy corresponding to the potential across the layer of 10–40 volts as measured by Mahl. Mahl interprets this to mean that there is field emission from the aluminum but that some electrons suffer energy losses in traversing the oxide. He could obtain the effect for bombarding voltages between 15 and 1000 volts. By momentarily cutting off the collector voltage and observing the partial neutralization of the surface charge, he obtained a value for the capacity between the surface and the aluminum base which agreed with the value calculated on the basis of the layer thickness.

Piore[81] has obtained Malter effect from surfaces prepared by evaporating barium borate or quartz on a metal plate and heat treating with cesium and oxygen. He obtained the same dependence of emission current on bombarding current and collector voltage as did Malter, although the particular constants involved changed from surface to surface. Bojinesco[115] used just Al_2O_3 on aluminum and obtained Malter effect with electron bombardment and also with bombardment by negative ions H^-, N^-, O^-, and O_2^-.

Using cathodes of the type described by Malter, Mühlenpfordt[102] has shown that on letting in argon or helium, at pressures of 10^{-5} mm. mercury or more, the thin film field emission discharge changed to a spray discharge just as described by Güntherschulze. He obtained two types of discharge: a low current discharge without cathodic scintillation and a high current discharge with scintillations. On pumping out the gas, he again obtained a Malter emission, the transition being a continuous one. He assumes that the surface charge is maintained by positive ions during the discharge and suggests that even under "high vacuum" conditions the contribution of such ions from the residual gas is not negligible but may play an essential role. In another experiment, Mühlenpfordt admitted oxygen at a pressure of 10^{-5} mm. mercury and then quickly pumped it out while the Malter current was decaying following cessation of primary bombardment. When the oxygen was admitted, the current dropped abruptly by three orders of magnitude and did not recover when the oxygen was pumped out. He had also observed that a freshly prepared surface must be activated by electron bombardment for some time before Malter effect occurred. On the basis of these observations, he suggested that a metallic cesium layer is essential to the process. However, when it is first evaporated onto the Al_2O_3, it comes into contact with the aluminum base through small holes in the oxide layer, thus effectively shorting it out. The subsequent oxidation of the cesium is designed to break up these short circuits and, finally, free cesium is produced only at the surface by the electron bombardment.

Zernov[240,241] believes that the role of the Cs_2O is to fill up the holes and irregularities in the Al_2O_3 layer rather than to provide free cesium eventually on the surface. He had also produced Malter effect with MgO films without the use of additional Cs_2O. He obtained a rapid decrease in emitted current, following cessation of bombardment, and then a stabilization of the current presumably due to the effect of residual gas. This latter effect was obtained with surfaces of [Al]-Al_2O_3-Cs_2O and MgO.

A point that has been brought out only by Mühlenpfordt is that all these characteristics are strongly dependent on the target temperature. The emission drops sharply with increasing temperature and simultaneously the decay rate of the decaying Malter current increases. This is explained by the positive temperature coefficient of conductivity of the Al_2O_3. At higher temperatures the increased leakage current prevents the surface from charging up as much and causes it to decay more rapidly after the excitation is cut off.

Paetow[139] has conducted some experiments which may have considerable bearing on this subject. He sprinkled an insulating powder with grain size less than 1 micron on a metallic cathode and with a gas pressure between 10^{-3} and 10^{-7} mm. mercury started a gas discharge with a few thousand volts on the anode. He reported that almost any gas would do and that many types of powder could be used, e.g., quartz, glass, MgO, Al_2O_3. On initiation of the gas discharge, the powder flew around explosively but left an active layer of fine powder on the cathode which stuck tightly to the metal. He then obtained two forms of the discharge: a steady low voltage form and an unsteady high voltage discharge accompanied by cathodic scintillation, i.e., spray discharge, and he likens these to the two forms of Malter current. The low voltage discharge could be maintained at pressures of less than 10^{-6} mm. mercury and between 10^{-2} and 10^{-6} mm. mercury the current is essentially independent of applied voltage. The maximum current density so obtained was 100 ma/cm.^{2}; above this it changed over to the high voltage form. Owing to sputtering the target life was a few weeks. The low voltage discharge exhibited time lags in its I/V characteristic although much less than in Malter effect. Paetow suggests that field omission may take place around the edges of the grains rather than through the grains themselves. Investigations of this type may lead to a more complete understanding of the mechanism involved in the Malter effect.

BIBLIOGRAPHY

1. Whiddington, R. *Proc. Roy. Soc.*, **A86**, 360 (1912).
2. Terrill, H. M. *Phys. Rev.*, **22**, 161 (1922).
3. Becker, A. *Ann. Phys. Lpz.*, **78**, 228 (1925).
4. Becker, A. *Ann. Phys. Lpz.*, **78**, 253 (1925).
5. Farnsworth, H. E. *Phys. Rev.*, **25**, 41 (1925).
6. Lange, H. *Jahrb. drahtl. Telegr.*, **26**, 38 (1925).
7. Petry, R. L. *Phys. Rev.*, **26**, 346 (1925).
8. Tellegren, B. D. H. *Physica*, **6**, 113 (1926).
9. Brinsmade, J. B. *Phys. Rev.*, **30**, 494 (1927).
10. Davisson, C. J. and Germer, L. H. *Phys. Rev.*, **30**, 705 (1927).
11. Hyatt, J. M. *Phys. Rev.*, **32**, 922 (1928).
12. Stehberger, K. H. *Ann. Phys. Lpz.*, **86**, 825 (1928).
13. Becker, A. *Ann. Phys. Lpz.*, **2**, 249 (1929).
14. Daene H. and Schmerwitz, G. *Z. Phys.*, **53**, 404 (1929).
15. Sixtus, K. *Ann. Phys. Lpz.*, **3**, 1017 (1929).
16. Bethe, H. *Ann. Phys. Lpz.*, **5**, 325 (1930).
17. Rao, S. R. *Proc. Roy. Soc.*, **A128**, 41, 57 (1930).
18. Richardson, O. W. *Proc. Roy. Soc.*, **A128**, 63 (1930).
19. Soller, T. *Phys. Rev.*, **36**, 1212 (1930).
20. Ahearn, A. J. *Phys. Rev.*, **38**, 1858 (1931).
21. Farnsworth, H. E. *Phys. Rev.*, **40**, 684 (1932).
22. Fröhlich, H. *Ann. Phys. Lpz.*, **13**, 229 (1932).

23. Güntherschulze, A. *Z. Phys.*, **86,** 778 (1933).
24. Güntherschulze, A. and Fricke, H. *Z. Phys.*, **86,** 451 (1933).
25. de Lussanet de la Sabloniere, C. J. *Hochfrequenztech. u. Elektroakust.*, **41,** 195 (1933).
26. Copeland, P. L. *Phys. Rev.*, **46,** 167 (1934).
27. Farnsworth, P. T. *J. Franklin Inst.*, **2,** 411 (1934).
28. Becker, J. A. *Rev. Mod. Phys.*, **7,** 95 (1935).
29. de Boer, J. H. Electron Emission and Adsorption Phenomena, Cambridge University Press, Cambridge, England, 1935.
30. Haworth, L. J. *Phys. Rev.*, **48,** 88 (1935).
31. Hayner, L. J. *Physics*, **6,** 323 (1935).
32. Iams, H. and Salzberg, B. *Proc. Inst. Radio Engrs.*, **23,** 55 (1935).
33. Langenwalter, H. W. *Ann. Phys. Lpz.*, **24,** 273 (1935).
34. Penning, F. M. and Kruithof, A. A. *Physica*, **2,** 793 (1935).
35. MacColl, L. A. *Phys. Rev.*, **56,** 699 (1939).
36. Schockley, W. *Phys. Rev.*, **56,** 317 (1939).
37. Nichols, M. H. *Phys. Rev.*, **57,** 297 (1940).
38. Was, D. A. and Tol, T. *Physica*, **7,** 253 (1940).

1936

39. Afanasjeva, A. V. and Timofeev, P. V. *Physik. Z. Sowjetunion*, **10,** 831. The Secondary Electron Emission from Oxidized Silver and Molybdenum Surfaces.
40. Afanasjeva, A. V., Timofeev, P. V. and Ignatov, A. S. *J. Tech. Phys. USSR*, **6,** 1649. On the Secondary Emission of Electrons from Thin Films Deposited on Glass.
41. Bruining, H. *Physica*, **3,** 1046. The Depth at Which Secondary Electrons Are Liberated.
42. Dobroljubski, A. N. *J. Tech. Phys. USSR*, **6,** 1489. On the Secondary Electron Emission from Composite Surfaces.
43. Dobroljubski, A. N. *Z. Phys.*, **102,** 626. On The Relation of Secondary Electron Emission to The Photoeffect and Thermionic Effect.
44. Dobroljubski, A. N. *Physik. Z. Sowjetunion*, **10,** 242. On the Correlation of The Secondary Emission of Electrodes Possessing Photosensitivity and the Thermoeffect of Ions.
45. Farnsworth, H. E. *Phys. Rev.*, **49,** 605. Penetration of Low Speed Diffracted Electrons.
46. Haworth, L. J. *Phys. Rev.*, **50,** 216. Energy Distribution of Secondary Electrons from Colombium.
47. Knoll, M. *Naturwissenschaften*, **24,** 345. Variations in the Secondary Electron Emission from Insulators and Semi-Conductors On Electron Bombardment.
48. Kwarzchawa, I. F. *Physik. Z. Sowjetunion*, **10,** 809. Secondary Emission And Fatigue Phenomena in Photosensitive Caesium-Oxygen Electrodes.
49. Malter, L. *Phys. Rev.*, **50,** 48. Thin Film Field Emission.
50. Rudberg, E. *Phys. Rev.*, **50,** 138. Inelastic Scattering of Electrons from Solids.
51. Rudberg, E. and Slater, J. C. *Phys. Rev.*, **50,** 150. Theory of Inelastic Scattering of Electrons in Solids.
52. Strübig, H. *Phys. Z.*, **37,** 402. The Potential of A Target Insulated in High Vacuum on Bombardment by Electrons.
53. Timofeev, P. V. and Pyatnitski, A. I. *Physik. Z. Sowjetunion*, **10,** 518. On the Secondary Emission of Electrons From Caesium-Oxygen Cathodes.

54. Treloar, L. R. G. *Nature, Lond.*, **137**, 579. Relation Between Secondary Emission and Work Function.
55. Treloar, L. R. G. *Proc. Phys. Soc. Lond.*, **48**, 488. A Method of Measuring Secondary Electron Emission from Filaments.
56. Warnecke, R. *J. phys. radium*, **7**, 270. Secondary Emission of Pure Metals.
57. Warnecke, R. *J. phys. radium*, **7**, 318. Critical Potentials of Secondary Emission.
58. Weiss, G. *Z. techn. Phys.*, **17**, 623. On the Secondary Electron Multiplier.
59. Ziegler, M. *Physica*, **3**, 1. Shot Effect of Secondary Emission. Part I.
60. Ziegler, M. *Physica*, **3**, 307. Shot Effect of Secondary Emission. Part II.
61. Zworykin, V. K., Morton, G. A. and Malter, L. *Proc. Inst. Radio Engrs.*, **24**, 351. The Secondary Emission Multiplier—A New Electronic Device.

1937

62. Afanasjeva, A. V. and Timofeev, P. V. *Tech. Phys. USSR*, **4**, 953. The Secondary Electron Emission of Gold, Silver and Platinum Covered with Thin Layers of the Alkali Metals.
63. Bhawalker, P. R. *Proc. Indian Acad. Sci.* **A6**, 74. An Explanation of the Maximum in Secondary Electron Emission of Metals.
64. Brown, J. B. *J. Opt. Soc. Amer.*, **27**, 186. Brightness of Cathode-Luminescence at Low Current and Low Voltages.
65. Bruining, H., de Boer, J. H. and Burgers, W. G. *Physica*, **4**, 267. Secondary Electron Emission in Valves With Oxide Cathodes.
66. Bruining, H. and de Boer, J. H. *Physica*, **4**, 473. Secondary Electron Emission of Metals With a Low Work Function.
67. Dobroljubski, A. N. *Physik. Z. Sowjetunion*, **11**, 118. Some Data on The Question of the Relation Between Secondary Electron Emission and Photosensitivity.
68. Hagen, C. and Bey, A. *Z. Phys.*, **104**, 681. The Charging Potential of Substances Bombarded by Electrons.
69. Herold, K. *Funktech. Monatshaft*, **9**, 271. Use of Secondary Electron Emission in High Frequency and Amplifier Technique.
70. Katz, H. *Z. techn. Phys.*, **18**, 555. Penetration of Slow Electrons Through Metal Foils.
71. Kluge, W., Beyer, O. and Steyskal, H. *Z. techn. Phys.*, **18**, 219. On Photoelectric Cells with Secondary Emission Amplification.
72. Kollath, R. *Phys. Z.*, **38**, 202. Secondary Electron Emission of Solids.
73. Koller, L. R. and Johnson, R. P. *Phys. Rev.*, **52**, 519. Visual Observations on the Malter Effect.
74. Krenzien, O. *Z. techn. Phys.*, **18**, 568. Production of Secondary Electrons in Adsorbed Layers.
75. Kubetzky, L. A. *Proc. Inst. Radio Engrs.*, **25**, 421. Multiple Amplifier.
76. Kurrelmeyer, B. and Hayner, L. J. *Phys. Rev.*, **52**, 952. Shot Effect of Secondary Emission from Nickel and Beryllium.
77. Mahl, H. *Z. techn. Phys.*, **18**, 559. Field Emission from Composite Cathodes with Electron Bombardment.
78. Müller, H. O. *Z. Phys.*, **104**, 475. Dependence of Secondary Emission Upon Primary Angle.
79. Myers, D. M. *Proc. Phys. Soc. Lond.*, **49**, 264. The Division of Primary Electron Current Between Grid and Anode of a Triode.

80. Nottingham, W. B. *J. Appl. Phys.*, **8**, 762. Electrical and Luminescent Properties of Willemite Under Electron Bombardment.
81. Piore, E. R. *Phys. Rev.*, **51**, 1111. Thin Film Field Emission.
82. Timofeev, P. V. and Pyatnitski, A. I. *Tech. Phys. USSR*, **4**, 945. Secondary Electron Emission from Complex Cathodes of Rubidium and Potassium.
83. Treloar, L. R.G. *Proc. Phys. Soc. Lond.*, **49**, 392. Secondary Electron Emission from Complex Surfaces.
84. Warnecke, R. *L'Onde Electr.*, **16**, 509. The Principal Laws of Secondary Electron Emission From Metallic Surfaces.

1938

85. Bay, Z. *Nature, Lond.*, **141**, 284, 1011. The Electron Multiplier As An Electron Counting Device.
86. Bruining, H. *Philips Tech. Rev.*, **3**, 80. Secondary Electron Emission.
87. Bruining, H. and de Boer, J. H. *Physica*, **5**, 17. Secondary Electron Emission Part I. Secondary Electron Emission of Metals.
88. Bruining, H. *Physica*, **5**, 901. Secondary Electron Emission Part II. Absorption of Secondary Electrons.
89. Bruining, H. *Physica*, **5**, 913. Secondary Electron Emission Part III. Secondary Electron Emission Caused by Bombardment with Slow Primary Electrons.
90. Copeland, P. L. *Phys. Rev.*, **53**, 328. Secondary Electron Emission from Sodium Films on Tantalum.
91. Jonker, J. L. H. and Teves, M. C. *Philips Tech. Rev.*, **3**, 133. Technical Applications of Secondary Emission.
92. Jonker, J. L. H. *Philips Tech. Rev.*, **3**, 211. Phenomena in Amplifier Valves Caused by Secondary Emission.
93. Jonker, J. L. H. and v. Overbeek, A. J. *Wireless Engr.*, **15**, 150. The Application of Secondary Emission in Amplifying Valves.
94. Khlebnikov, N. S. and Korshunova, A. *Tech. Phys. USSR*, **5**, 363. The Secondary Emission of Composite Surfaces.
95. Khlebnikov, N. S. *Tech. Phys. USSR*, **5**, 593. The Influence of Gases on the Secondary Emission of Certain Metals.
96. Kollath, R. *Naturwissenschaften*, **26**, 60. The Influence of the Geometrical Arrangement of Atoms on Secondary Electron Emission.
97. Kollath, R. *Z. techn. Phys.*, **19**, 602. A Secondary Electron Emission Experiment.
98. Kollath, R. *Ann. Phys. Lpz.*, **33**, 285. On The Secondary Electron Emission of Beryllium.
99. Lukjanov, S. J. *Physik. Z. Sowjetunion*, **13**, 123. Variation of Secondary Emission Yield with Angle of Incidence.
100. Mahl, H. *Z. techn. Phys.*, **19**, 313. Field Emission from Composite Cathodes with Electron Bombardment. Part II.
101. Morgulis, N. D. and Nagorsky, A. *Tech. Phys. USSR*, **5**, 848. Secondary Electron Emission from Oxide Coated Cathodes.
102. Muhlenpfort, J. *Z. Phys.*, **108**, 698. Electron Field Emission from Thin Insulating Layers of the Type Al-Al_2O_3-Cs_2O.
103. Nelson, H. *J. Appl. Phys.*, **9**, 592. Method of Measuring Luminescent Screen Potential.
104. Nelson, H. *Phys. Rev.*, **55**, 985. Phenomenon of Secondary Electron Emission.
105. Sandhagen, M. *Z. Phys.*, **110**, 553. Measurements of Secondary Electrons Arising from Reflector Grids.

105a. Pyatnitsky, A., *J. Tech. Phys. USSR*, **8,** 1014. Distribution of the Energy of Secondary Electrons Emitted by a Composite Cesium Cathode.

106. Schneider, E. G. *Phys. Rev.*, **54,** 185. Secondary Emission of Beryllium.

107. Shockley, W. and Pierce, J. R. *Proc. Inst. Radio Engrs.*, **26,** 321. A Theory of Noise for Electron Multipliers.

108. Treloar, L. R. G. and Landon, D. H. *Proc. Phys. Soc. Lond.*, **50,** 625. Secondary Electron Emission from Nickel, Cobalt and Iron as a Function of Temperature.

109. Treloar, L. R. G. *Wireless Engr.*, **15,** 535. The Measurement of Secondary Emission in Valves.

110. Turnbull, J. C. and Farnsworth, H. E. *Phys. Rev.*, **54,** 509. Inelastic Scattering of Slow Electrons from a Silver Single Crystal.

111. Vudynski, M. *J. Tech. Phys. USSR*, **8,** 790. The Investigation of Secondary Electron Emission from Dielectrics by a Thermal Method.

112. Warnecke, R. and Lortie, M. *J. phys. radium*, **9,** Suppl 8. Relation Between the Coefficient of Secondary Emission and the Work Function of Metal Surfaces.

1939

113. Allen, J. S. *Phys. Rev.*, **55,** 966. The Detection of Single Positive Ions, Electrons and Photons by a Secondary Emission Multiplier.

114. Bojinesco, A. *C. R. Acad. Sci. Paris*, **209,** 512. Energy Distribution of Low Temperature Secondary Electrons.

115. Bojinesco, A. *C. R. Acad. Sci. Paris*, **209,** 1800. Electronic Field Emission After Bombardment of Aluminum Oxide by Electrons or Negative Ions H^-, N^-, O^-, O_2^-.

116. Borzyak, P. *J. Tech. Phys. USSR*, **15,** 1380. The Emission from Composite Cathodes under Simultaneous Electron Bombardment and Illumination

117. Borzyak, P. *J. Tech. Phys. USSR*, **15,** 2032. Relation between The Emission Characteristics and The Conductivity of Oxide Caesium Photocathodes

118. De Boer, J. H. and Bruining, H. *Physica*, **6,** 941. Secondary Electron Emission. Part VI. The Influence of Externally Adsorbed Ions and Atoms on the Secondary Electron Emission of Metals.

119. Bruining H. and de Boer, J. H. *Physica*, **6,** 823. Secondary Electron Emission Part IV. Compounds with a High Capacity for Secondary Electron Emission.

120. Bruining, H. and de Boer, J. H. *Physica*, **6,** 834. Secondary Electron Emission. Part V. The Mechanism of Secondary Electron Emission.

121. Coomes, E. A. *Phys. Rev.*, **55,** 519. Total Secondary Electron Emission from Tungsten and Thorium Coated Tungsten.

122. Copeland, P. L. *Phys. Rev.*, **55,** 1270. Secondary Electron Emission from Sodium Films Contaminated with Gas.

123. von Frerichs, R. and Krautz, E. *Phys. Z.*, **40,** 229. A Simple Arrangement For The Measurement of the Charging Potential of Phosphor Layers Bombarded by Electrons.

124. Hagen, C. *Phys. Z.*, **40,** 621. Charging Potential, Secondary Electron Emission and Fatigue Phenomena in Metals and Phosphors Bombarded by Electrons.

125. Hintenberger, H. *Z. Phys.*, **114,** 98. On Secondary Emission and Charging Phenomena of Insulators.

126. Jonker, J. L. H. *Wireless Engr.*, **16,** 274. Pentode and Tetrode Output Valves.

127. Khlebnikov, N. S. *J. Tech. Phys. USSR*, **9**, 367. Certain Properties of Effective Secondary Electron Emitters.
128. Knoll, M. and Theile, R. *Z. Phys.*, **113**, 260. Electron Pictures of The Structure of Surfaces and Thin Layers.
129. Korshunova, A. S. and Khlebnikov, N. S. *J. Tech. Phys. USSR*, **9**, 860. Secondary Electron Emission from Thin Dielectric Layers.
130. Kosman, M., Abramov, A. and Gurilev, B. *J. Exp. Theoret. Phys. USSR*, **9**, 176. Secondary Electron Emission of Mica.
131. Krautz, E. *Z. Phys.*, **114**, 459. On Charging and The Reduction of Charging of Phosphors and Semi-Conductors Bombarded by Electrons.
132. Kushnir, Yu. M. and Milyutin, I. *J. Tech. Phys. USSR*, **9**, 267. Secondary Electron Emission Under the Action of Two Electron Beams.
133. Kushnir, Yu. M. and Milyutin, I. *J. Tech. Phys. USSR*, **9**, 1589. On The Secondary Electron Emission from Mercury.
134. Mahl, H. *Jahrb. AEG-Forschung*, **6**, 33. Observations of the Secondary Electron Emission of Evaporated Alkali Layers by an Oscillographic Method.
135. Majewski, W. *Acta. Phys. Polon.*, **7**, 327. Contribution to The Measurement Technique of Secondary Electron Emission.
136. Martin, S. T. and Headrick, L. B. *J. Appl. Phys.*, **10**, 116. Light Output and Secondary Emission Characteristics of Luminescent Materials.
137. Morgulis, N. D. *J. Tech. Phys. USSR*, **9**, 853. Nature of Secondary Electron Emission from Composite Cathodes.
138. Nelson, H. *Phys. Rev.*, **55**, 985. Phenomenon of Secondary Electron Emission.
139. Paetow, H. *Z. Phys.*, **111**, 770. On the Effect of a Gas Discharge on an Electrode Emitting Electrons and the Field Emission from Thin Insulating Layers.
140. Pes'yatski, I. F. *J. Tech. Phys. USSR*, **9**, 188 Secondary Electron Emission from Thin Films.
141. Rakov, V. I. and Antonov, V. A. *J. Tech. Phys. USSR*, **9**, 870. Secondary Electron Emission of Tungsten, Copper, Iron at High Voltages.
142. Rann, W. H. *J. Sci. Instrum.*, **16**, 241. Amplification by Secondary Electron Emission.
143. Saegusa, H. and Matsumoto, T. *Tohoku Imp. Univ.; Science Repts.*, **28**, 245. Total Secondary Electron Emission from Nickel, Sodium Chloride and Potassium Chloride.
144. Suhrmann, R. and Kundt, W. *Naturwissenschaften*, **27**, 548. The Secondary Electron Emission of Clean Metals in the Unordered and Ordered Condition.
145. Suhrmann, R. and Kundt, W. *Naturwissenschaften*, **27**, 707. Concerning the Mechanism of Secondary Electron Emission.
146. Timofeev, P. V. *C. R. Acad. Sci. URSS*, **25**, 11. Mechanism of Secondary Electron Emission from Composite Surfaces.
147. Vudynski, M. M. *J. Tech. Phys. USSR*, **9**, 271. Secondary Electron Emission from Thin Dielectric Layers.
148. Vudynski, M. M. *J. Tech. Phys. USSR*, **9**, 1377. On The Nature of Particles Emitted from Sodium Chloride when Bombarded by Electrons.
149. Vudynski, M. M. *J. Tech. Phys. USSR*, **9**, 1583. On the Velocity Distribution of Secondary Electrons Emitted from Sodium Chloride.
150. Warnecke, R. and Lortie, M. *C. R. Acad. Sci. Paris*, **208**, 429. On the Secondary Emission of Beryllium.
151. Wooldridge, D. E. *Phys. Rev.*, **56**, 562. Theory of Secondary Emission.
152. Wooldridge, D. E. *Phys. Rev.*, **56**, 1062. The Secondary Electron Emission from Evaporated Nickel and Cobalt.

153. Yasnopol'ski, N. and Tyagunov, G. A. *J. Tech. Phys. USSR*, **9**, 1573. On Secondary Electron Emission.

154. Zworykin, V. K. and J. A. Rajchman. *Proc. Inst. Radio Engrs.*, **27**, 558. The Electrostatic Electron Multiplier.

1940

155. Copeland, P. L. *Phys. Rev.*, **58**, 604. Secondary Emission from Films of Platinum on Aluminum.

156. Frimer, A. I. *J. Tech. Phys. USSR*, **5**, 394. A Study of the Secondary Electron Emission from Copper Oxide, Pure or Treated with Alkali Metals.

157. Gubanov, A. *J. Exp. Theoret. Phys. USSR*, **2**, 161. The Effect of a Charge on an Electron Beam during Secondary Emission.

158. Hastings, A. E. *Phys. Rev.*, **57**, 695. Secondary Emission from Films of Silver on Platinum.

159. Heimann, W. and Geyer, K. *Elekt. Nachr.-Tech.*, **17**, 1. Direct Measurement of the Secondary Electron Yield from Insulators.

160. Kadyschevitsch, A. E. *J. Phys. USSR*, **2**, 115. Theory of Secondary Electron Emission from Metals.

161. Kamogawa, H. *Phys. Rev.*, **58**, 660. Secondary Emission and Electron Diffraction on the Glass Surface.

162. Kirvalidze, I. D. *C. R. Acad. Sci. URSS*, **26**, 635. A Method for Determining the Charging Potential of Dielectrics and the Lower Limit of Secondary Electron Emission from a Monocrystal of NaCl.

163. Kollath, R. *Z. techn. Phys.*, **21**, 328. A New Method For the Measurement of the Energy Distribution of Secondary Electrons.

164. Morgulis, N. D. *J. Tech. Phys. USSR*, **10**, 79. The Mechanism of Secondary Electron Emission from Composite Surfaces.

165. Morgulis, N. D. *J. Tech. Phys. USSR*, **10**, 1710. On the Nature of Secondary Emission from Composite Cathodes.

166. Morgulis, N. D. and Dyatlovitskaya, B. I. *J. Tech. Phys. USSR*, **10**, 657. On the Emission from Antimony-Caesium Cathodes.

167. Nelson, H. *Phys. Rev.*, **57**, 560. Field Enhanced Secondary Electron Emission.

168. Piore, E. R. and Morton, G. A. *J. Appl. Phys.*, **11**, 153. The Behavior of Willemite Under Electron Bombardment.

169. Reichelt, W. *Ann. Phys. Lpz.*, **38**, 293. Influence of Temperature on The Secondary Emission of Metals.

170. Salow, H. *Z. techn. Phys.*, **21**, 8. On The Secondary Electron Yield of Electron Bombarded Insulators.

171. Salow, H. *Phys. Z.*, **41**, 434. Angular Dependence of The Secondary Electron Emission from Insulators.

172. Scherer, K. *Arch. Elektrotech.*, **34**, 143. Charging and Secondary Electron Emission.

173. Schnitger, H. *Z. techn. Phys.*, **21**, 376. The Properties of Secondary Emitting Layers of Magnesium Oxide.

174. Teves, M. C. *Philips Tech. Rev.*, **5**, 253. A Photocell with Amplification By Means of Secondary Emission.

175. Timofeev, P. V. and Afanasjeva, A. V. *J. Tech. Phys. USSR*, **10**, 28. Secondary Emission from Oxides of Metals.

176. Timofeev, P. V. and Aranovich, P. M. *J. Tech. Phys. USSR*, **10**, 32. Barium Oxide and Magnesium Oxide Emitters of Secondary Electrons.

177. Timofeev, P. V. and Lunkova, J. *J. Tech. Phys. USSR*, **10**, 12. Electron Emission From Caesium Oxide Cathodes with Gold Particles in the Intermediate Layer.

178. Timofeev, P. V. and Lunkova, J. *J. Tech. Phys. USSR*, **10**, 20. Antimony-Caesium Emitters.

179. Timofeev, P. V. and Pyatnitski, A. I. *J. Tech. Phys. USSR*, **10**, 39. Secondary Electron Emission from Oxygen-Caesium Emitters at Different Primary Current Densities.

180. Timofeev, P. V. and Yumatov, K. A. *J. Tech. Phys. USSR*, **10**, 8. Secondary Electron Emission from Oxygen-Caesium Emitters at Low Velocities of Primary Electrons.

181. Timofeev, P. V. and Yumatov, K. A. *J. Tech. Phys. USSR*, **10**, 24. Secondary Electron Emission from Sulphur-Caesium Emitters.

182. Varadachari, P. S. *Proc. Indian Acad. Sci.*, **A12**, 381. Secondary Electron Emission of Nickel at the Curie Point.

183. Wooldridge, D. E. *Phys. Rev.*, **57**, 1080. Temperature Effects on the Secondary Electron Emission from Pure Metals.

184. Wooldridge, D. E. *Phys. Rev.*, **58**, 316. Temperature Effects In Secondary Emission.

185. Wooldridge, D. E. and Hartman, C. D. *Phys. Rev.*, **58**, 381. The Effects of Order and Disorder on Secondary Electron Emission.

186. Yasnopol'ski, N. *J. Tech. Phys. USSR*, **10**, 1813. On the "Jumps" Observed in Emitters of Poor Conductivity Caused by the Blocking Effect.

1941

187. Bay, Z. *Rev. Sci. Instrum.*, **12**, 127. Electron Multiplier as an Electron Counting Device.

188. Bekow, G. *Phys. Z.*, **42**, 144. Secondary Emission from Copper Single Crystals at Small Primary Velocities.

189. Bethe, H. A. *Phys. Rev.*, **59**, 940. On the Theory of Secondary Emission.

190. Bruining, H. *Physica*, **8**, 1161. Secondary Electron Emission from Metals with Low Work Function.

191. Chaudri, R. M. and Khan, A. W. *Phil. Mag.*, **31**, 382. Secondary Electron Emission from Nickel.

192. Friedheim, J. and Weiss, J. G. *Naturwissenschaften*, **29**, 777. Secondary Electron Yield of Silver-Magnesium Alloys.

193. Gille, G. *Z. techn. Phys.*, **22**, 228. Secondary Electron Emission by a Nickel-Beryllium Alloy.

194. Görlich, P. *Phys. Z.*, **42**, 129. Contribution To The Problem of Secondary Electron Emission of Condensed Alkali-Earth Metal Films.

195. Joffe, M. S. and Nechaev, I. V. *J. Exp. Theoret. Phys., USSR*, **11**, 93. The Secondary Electron Emission from Potassium.

196. Kadyschevitsch, A. E. *J. Phys. USSR*, **4**, 341. Theory of Secondary Electron Emission from Dielectrics and Semiconductors.

197. Kollath, R. *Ann. Phys. Lpz.*, **39**, 19. On the Influence of Temperature on the Secondary Electron Emission of Metals.

198. Kollath, R. *Ann. Phys. Lpz.*, **39**, 59. On the Energy Distribution of Secondary Electrons.

199. Kushnir, Yu. M. and Frumin, M. I. *J. Tech. Phys. USSR*, **11**, 317. The Dependence of the Energy Distribution Function of Secondary Electrons on the Angle of Emergence.

200. Malter, L. *Proc. Inst. Radio Engrs.*, **29**, 587. The Behavior of Electrostatic Electron Multipliers as a Function of Frequency.

201. Mathes, I. *Z. techn. Phys.*, **22**, 232. Secondary Electron Emission Properties of Some Alloys.

202. Maurer, G. *Z. Phys.*, **118**, 122. The Secondary Electron Emission From Semiconductors and Insulators.

203. Mishibori, E. *Proc. Phys.-Math. Soc. Japan*, **23**, 570. Secondary Electron Emission from Magnesium Oxide.

204. Morgulis, N. D. *Bull. Acad. Sci. URSS*, **5**, 536. The Emission of Electrons by Active Semi-Conducting Surfaces.

205. Morozov, P. M. *J. Exp. Theoret. Phys. USSR*, **11**, 402. The Effect of Temperature on Secondary Electron Emission.

206. Morozov, P. M. *J. Exp. Theoret. Phys. USSR*, **11**, 410. Secondary Electron Emission from Lead, Tin and Bismuth in the Solid and Liquid State.

207. Nemilov, Yu. A. *J. Tech. Phys. USSR*, **11**, 854. A New Method for Studying Secondary Electron Emission.

208. Paetow, H. *Z. Phys.*, **117**, 399. A New Form of Field Emission at Very Low Pressures from Metallic Surfaces Covered by a Deposit of Insulating Material.

209. Randenbusch, H. *Z. Tech. Phys.* **22**, 237. Some Investigations on the Technical Uses of Secondary Emission Surfaces.

210. Tanaka, M. *Proc. Phys.-Math. Soc. Japan*, **22**, 899. After Effect of Metal Bombarded by Electrons.

211. Teichmann, H. and Geyer, K. *Z. ges. Naturw.*, **7**, 313. The Occurrence of Structure Dependent Selectivity in Secondary Electron Emission.

212. Vudynski, M. M. *J. Tech. Phys USSR*, **11**, 1066. The Stability of Secondary Electron Emission from Alkali Halide Cathodes.

213. Wecker, F. *Ann. Phys. Lpz.*, **40**, 405. New Measurements on the Absorption, Back Diffusion and Secondary Emission in Aluminum and Gold.

214. Wolff, H. *Ann. Phys. Lpz.*, **39**, 591. Secondary Electron and Photoelectron Emitting Semi-Conductors.

215. Zworykin, V. K. Ruedy, J. E. and Pike, E. W. *J. Appl. Phys.*, **12**, 696. Silver-Magnesium Alloy as a Secondary Emitting Material.

1942

216. Bruining, H. The Secondary Electron Emission of Solids, Springer, Berlin.

217. Geyer, K. H. *Ann. Phys. Lpz.*, **41**, 117. On the Properties of Yield and Energy Distribution of Secondary Electrons from Evaporated Layers of Increasing Thickness.

218. Geyer, K. H. *Ann. Phys. Lpz.*, **42**, 241. Observations on the Secondary Electron Emission from Nonconductors.

219. Görlich, P. *Phys. Z.*, **43**, 121. On the Secondary Emission of Evaporated Layers of Antimony.

220. Krenzien, O. *Wiss. Veroff. Siemens-Werk.*, **20**, 91. The Elementary Processes in the Secondary Electron Emission of Polar Crystals.

221. McKay, K. G. *Phys. Rev.*, **61**, 708. Total Secondary Emission from Thin Films of Sodium on Tungsten.

222. Skellett, A. M. *J. Appl. Phys.*, **13**, 519. Use of Secondary Electron Emission to Obtain Trigger or Relay Action.

223. Truell, R. *Phys. Rev.*, **62**, 340. Range of Secondary Electrons in Magnesium.

1943

224. Bronstein. *J. Tech. Phys. USSR*, **13**, 176.
225. Geyer, K. H. *Ann. Phys. Lpz.*, **42**, 337. Contribution to our Knowledge of the Fundamental Process of Secondary Electron Emission.
226. Gimpel, I. and Richardson, O. *Proc. Roy. Soc.*, **A182**, 17. The Secondary Electron Emission from Metals in the Low Primary Energy Region.
227. Kennedy, W. R. and Copeland, P. L. *Phys. Rev.*, **63**, 61. The Temperature Coefficient of the Secondary Emission Yield and the Work Function of Molybdenum Coated with Beryllium and Treated with HCl.
228. Schlectweg, H. *Naturwissenschaften*, **31**, 204. Quantum Theory of Secondary Electron Emission of Transition Metals.
229. Suhrmann, R. and Kundt, W. *Z. Phys.*, **120**, 363. The Secondary Emission of Pure Metallic Films in the Ordered and Unordered States and Their Transparency for Secondary Electrons.
230. Suhrmann, R. and Kundt, W. *Z. Phys.*, **121**, 118. On the Effect of Adsorbed Oxygen on the Secondary Emission of Vaporized Metal Films at 293°K and 83°K.
231. Trey, F. *Phys. Z.*, **44**, 38. Review on the Effect of Field Emission on Secondary Electron Emission.

1944

232. Afanasjeva, A. V. *Univ. M. V. Lomousova*, **74**, 114. Stable Emitters of Secondary Electrons.
233. Aranovich, R. M. *Bull. Acad. Sci. URSS*, **8**, 346. Electronic Devices with Effective Emitters of Secondary Electrons.
234. Harries, J. H. O. *Electronics*, **10**, 100, (Sept). Secondary Electron Radiation.
235. Johnson, J. B. *Phys. Rev.*, **66**, 352. Enhanced Thermionic Emission.
236. Knoll, M., Hachenberg, O. and Randmer, J. *Z. Phys.*, **122**, 137. Mechanism of Secondary Emission in the Interior of Ionic Crystals.
237. Kubetskii, L. A. *Bull. Acad. Sci. URSS*, **8**, 357. Some Results of the Use of Secondary Emission Multipliers.
238. Kwarzchawa, I. F. *Bull. Acad. Sci. URSS*, **8**, 373. Change of Conductivity of Aluminum Oxide upon Electron Bombardment.
239. Lukjanov, S. J. *Bull. Acad. Sci. URSS*, **8**, 330. Secondary Electron Emission of Solids.
240. Zernov, D. V., Elinson, M. I. and Levin, N. M. *Bull. Acad. Sci. URSS, Classe sci. tech.* 166. Investigation of Autoelectronic Emission of Thin Dielectric Films.
241. Zernov, D. V. *Bull. Acad. Sci. URSS*, **8**, 352. On the Influence of Strong Electric Fields on the Secondary Electron Emission of Dielectric Films.

1945

242. Kadyschevitsch, A. E. *J. Phys., USSR*, **9**, 431. The Velocity Distribution of Secondary Electrons of Various Emitters.
243. Kadyschevitsch, A. E. *J. Phys., USSR*, **9**, 436. On the Measurement of the Depth of Generation of Secondary Electrons in Metals.
244. Mueller, C. W. *J. Appl. Phys.*, **16**, 453. The Secondary Electron Emission of Pyrex Glass.

245. Sorg, H. E. and Becker, G. A. *Electronics*, **18**, July, 104. Grid Emission in Vacuum Tubes.

246. Timofeev, P. V. *Bull. Acad. Sci. URSS*, **8**, 340. The Role of Surface Charges in Electronic Devices.

247. Wang, C. C. *Phys. Rev.*, **68**, 284. Reflex Oscillator Utilizing Secondary Emission Currents.

1946

248. Johnson, J. B. *Phys. Rev.*, **69**, 693. Secondary Emission of Thermionic Oxide Cathodes.

249. Johnson, J. B. *Phys. Rev.*, **69**, 702. Enhanced Thermionic Emission from Oxide Cathodes.

250. Koller, L. R. and Burgess, J. S. *Phys. Rev.*, **70**, 571. Secondary Emission from Germanium, Boron and Silicon.

251. Lallemand, A. *Rev. Sci. Paris*, **84**, 131. Application of Secondary Emission of Electrons to Multiplier Tubes.

252. Pomerantz, M. A. *J. Franklin Inst.*, **241**, 415; **242**, 41. Secondary Electron Emission from Oxide Coated Cathodes.

253. Pomerantz, M. A. *Phys. Rev.*, **70**, 33. Temperature Dependence of Secondary Electron Emission from Oxide Coated Cathodes.

254. Sard, R. D. *J. Appl. Phys.*, **17**, 768. Calculated Frequency Spectrum of the Shot Noise from a Photo-Multiplier.

1947

255. Frimer, A. I. *J. Tech. Phys. USSR*, **17**, 71. Study of Secondary Emission at Low Temperatures.

256. Greenblatt, M. H. and Miller, P. A., Jr. *Phys. Rev.*, **72**, 160. A Microwave Secondary Emission Multipler.

257. Johnson, J. B. *Phys. Rev.*, **73**, 1058 (1948). Secondary Electron Emission from Targets of Barium-Strontium Oxide.

258. Mendenhall, H. E. *Phys. Rev.*, **72**, 532. Secondary Emission from Conducting Films of Tin Oxide.

258a. Palluel, P. *C. R. Acad. Sci. Paris*, **224**, 1492. Rediffused Component of Secondary Radiation from Metals.

258b. Palluel, P. *C. R Acad. Sci. Paris*, **224**, 1551. On the Mechanism of Rediffusion of Electrons from Metals.

258c. Palluel, P. *C. R. Acad. Sci. Paris*, **225**, 383. The True Secondary Emission Coefficient in Metals.

259. Trump, J. G. and Van de Graaf, R. J. *J. Appl. Phys.*, **18**, 327. Insulation of High Voltages in Vacuum.

260. Zernov, D. V. and Kuljvarskaya, B. S. *J. Tech. Phys. USSR*, **17**, 309. Investigations of the Temperature Dependence of the Electronic Emission of Dielectric Films Under the Influence of the Field of Positive Surface Charge.

Television Pickup Tubes and the Problem of Vision

A. ROSE

RCA Laboratories Division, Princeton, N. J.

CONTENTS

		Page
I.	Introduction	131
II.	Major Types of Pickup Devices .	132
III.	Number and Variety of Television Pickup Tubes	133
IV.	Comparison of Actual and Possible Pickup Tubes	134
V.	Ideal Performance	135
VI.	An Experimental Realization of Ideal Performance	141
VII.	Performance of Selected Pickup Devices	146
	1. Human Eye	146
	2. Photographic Film .	148
	3. Television Pickup Tubes	150
	a. Image Dissector	150
	b. Iconoscope	151
	c. Image Iconoscope	153
	d. Orthicon	154
	e. Image Orthicon	155
	4. Discussion of Performance Curves	157
VIII.	A Criterion for Noise Visibility	160
IX.	Intelligence vs. Bandwidth and Signal-to-Noise Ratio	163
X	Concluding Remarks	165
	References	165

I. INTRODUCTION

The visual process in its most refined state is a simple counting process. For a given exposure time, scene brightness, and optical system, a finite number of light quanta will be radiated from the scene and absorbed by the particular device that looks at the scene. The device counts these quanta and from their number and distribution determines how many discrete bits of information it can furnish. The arithmetic used to convert numbers of quanta into numbers of bits of information will be outlined in a later section. For the present an attempt will be made to contrast the almost endless variety and complexity of picture pickup devices with the simplicity and universality of the performance scale according to which any of these devices may be judged. Throughout this paper emphasis is placed on performance limited by the finite number of available light quanta. This is patently a fundamental limitation. Performance limited by electron optical or structural defects will receive only passing mention. The technical problems involved in removing these defects may indeed be the most important and the most difficult for the success of a particular device. They are, nevertheless, removable defects and not fundamental limitations.

II. MAJOR TYPES OF PICKUP DEVICES

To name television pickup tubes, photographic film, and the human eye is to name examples of the three major types of visual processes; electrical, chemical, and biological. The detailed mechanics of the operation of a television pickup tube has little in common with that of photographic film and even less, perhaps, in common with that of the human eye. This, in spite of literary references to television pickup tubes as "electric eyes" and in spite of pictorial explanations of the human eye confined to the parallelism of parts—lens, black box, and film—of an ordinary camera. What *is* common to all of these devices, especially if they are well designed, is some means for counting incoming light quanta. The particular means for, or mechanics of, counting may be as varied as the imagination of man and nature combined. Television tubes generally convert light quanta into photoelectrons and count the number of electrons by measuring a current or a voltage. Photographic film, by an intermediate chemical process, converts each quantum (or at most a small number of quanta) into an opaque granule of silver and observes the density of the silver deposit as a measure of the number of granules. The human eye, also, by an obscure intermediate process, converts a small number of incoming quanta into a single nerve pulse. The brain estimates the number of original quanta by the frequency of arrival of these nerve pulses and (a speculative thought) by the regularity of their arrival.

A television pickup tube differs again from the photographic process and from human vision by virtue of having to arrange its information (or picture) in a form convenient for transmission to a remote point. The "arranging" is done by a scanning process in which all of the bits of information in any one picture are strung end-to-end and passed single file through the television transmitter. At the television receiver, a converse arrangement, also a scanning process, accepts the bits of information and distributes them one at a time into their proper places to re-form the picture. Each picture is composed of a few hundred thousand bits of information or picture elements. And 30 pictures are transmitted/second. Because the persistence of vision of the eye is a few tenths of a second, neither the point-wise assembly of any one picture nor the discrete succession of pictures is resolved. Rather, the eye sees a smooth flow of events as in the original scene.

The scanning process dictates its own special problems and in some ways complicates the design of television pickup tubes. The scanning process does *not*, however, affect the performance *attainable* by these tubes as compared with the performance attainable by devices that do not use scanning.

III. NUMBER AND VARIETY OF TELEVISION PICKUP TUBES

Human vision, the photographic process and television not only compose the three broad approaches to the problem of seeing but also point up the variety of solutions to this problem. An even greater variety, numerically at least, is found within the television approach itself. Here, the patent literature offers literally hundreds of examples of ingenuity of design applied to television pickup tubes. Many more examples can be added to these by purely engineering combinations of the already patented devices.

With no intent of making the formula exhaustive, a large number of pickup tubes may be generated in the following fashion. Trace the steps in the formation of a picture. To each step assign a number equal to the number of ways that step may be performed. Take the product of all of the step numbers. The product, so formed, is an upper limit to the number of different pickup tubes that may be designed. Because not all of the ways of performing each step are known, this operation is more likely to yield a lower limit than an upper limit. The operation will be carried out for its suggestive value at this point. The coverage will be sufficient only for the purpose of illustration.

The steps in the formation of a picture are these. Incoming light quanta are converted into something that can be stored and counted. An element, usually called a target or mosaic, is provided capable of storing the photoproducts for a thirtieth of a second. The target is scanned by some means that can "count" the photoproducts at each point on the target. The scanning process may return the target to its previous unexposed state or not. If it does not, a separate erasing mechanism is needed so that cyclic operation is possible. The "count" made by the scanning means is passed on to an amplifier and through other circuits to the television transmitter. The amplifier has a background "count" or noise of its own which tends to obscure the "count" that is fed into it. For this reason it is highly desirable to have some noiseless amplifier within the pickup tube that can magnify its "count" to a conveniently high level before it is fed into the external amplifier. Five steps have just been enumerated: Conversion, storage, scanning or counting, erasure, and noiseless amplification. Illustrative examples of ways of performing each step will be cited.

Light quanta may be converted into photoelectrons, excited electrons, ions, excited atoms, or dissociated molecules. The photoelectrons in turn may be accelerated and focused on to a target to excite other electrons or to dissociate or ionize molecules. The target used for storage may be one sided or two sided, insulating or semiconducting, continuous or apertured. The target may be scanned by a beam of electrons or a beam of light. The electron scanning beam may be of the low, medium or high velocity types, each velocity range being qualitatively distinct. The scanning beam may simply strip off the charge pattern deposited on the target by the picture, or it may charge the target to some arbitrary uniform potential. The scanning beam may not even strike the target, in which case only the paths of the electrons in the scanning beam are controlled by the charge pattern. If the scanning process has not restored the target to the unexposed state, a separate means such as a steady electrical leakage, a uniform electron spray, or a second scanning beam may serve the purpose. Finally, the scanning beam, after it has been influenced by the target, can be amplified, substantially without distortion, by an electron multiplier. (Alternatively, electron multiplication could have taken place in the previous process of laying down the charge pattern on the target.) In those cases where the beam does not strike the target, but is only controlled by it, large scanning beam currents may be used that do not require further amplification before being fed into the external amplifier. Without actually assigning the previously mentioned step numbers, one can see that even this rapid and incomplete enumeration will lead to hundreds of different pickup tubes.

If the above recital of ways of making pickup tubes has been more confusing than informative, it was not without plan. The present paper in no way aims at being a manual for designing pickup tubes nor does it attempt a critical survey of the various proposals that have been made or could be made for these tubes. Such a paper would indeed have an extraordinarily small "public." The recital was introduced to underline the highly un-unique character of any one pickup tube—and there are scores of them. It was introduced as a self-evident argument for singling out only a few pickup tubes for close attention and these for the one purpose of tracing the successive realized approximations to ideal pickup tube performance. The chronological order turns out to have also a certain logical sequence.

IV. COMPARISON OF ACTUAL AND POSSIBLE PICKUP TUBES

In the opening sentence of this paper the visual process was described as a simple counting process. Conceptually it is easy to demonstrate both that the maximum, or ideal, performance is achieved insofar as each incoming quantum can be counted, and that such a count bears a simple relation to the intelligence that can be transmitted. Conceptually, also, one can design a simple pickup tube to satisfy this definition of ideal performance. It is perhaps something of a disappointment to find that the pickup tubes that have come into use are generally not simple. Either their construction is elaborate or their detailed operation difficult to analyze. What is more, they have still not exhausted the possibilities of ideal performance. This situation is not unusual in the early stages of an art, and television has yet to be matured in the atmosphere of a successfully commercialized system—that the first attempts are not only imperfect but elaborately imperfect. Another circumstance that restricts the freedom with which promising designs for pickup tubes can be plucked out of a file and converted into useful instruments is the number of side conditions such tubes must satisfy. These conditions apply largely to the target. Uniformity of response, for example, is one of the most difficult to satisfy. It is the problem of getting a hundred thousand little elements to act alike. Another side condition is that remnants of pictures shall not be noticeable in succeeding pictures, that is, that the process of erasure be substantially complete. Resolution in excess of the circuits is an obvious but not an easily met condition. Freedom from spurious signals is a condition that has ruled out many otherwise promising experimental tubes. Finally, there are reasonable limits to the kinds of targets that can be fabricated, at least without extensive and costly research. And the paper designs for targets are well ahead of the art of making them.

These considerations are meant not to justify the complexity or imperfection of present tubes but rather to give a measure of the difficulties of realizing the simplicity and perfection inherently possible in pickup tube design. The simple counting process that forms the basis for ideal performance will be outlined in the next two sections both analytically and experimentally.

V. Ideal Performance

The object of this section is twofold. First, a completely general relation will be derived for the amount of information a picture pickup device can furnish as a function of the brightness of scene at which it is directed. Ideal performance will be insured in this derivation by the assumption that the pickup device can count each absorbed quantum of light. Second, in so doing, a general and absolute scale of performance will be set up according to which the performance of actual pickup devices may be judged. Use will be made of this scale of performance in a later section.

Consider a square element of area of side length "h" on the photosensitive surface of the picture pickup device. Let this element of area absorb on the average N quanta in the exposure time alloted. Because the absorption of quanta is a random process, the average absorption number, N, will have associated with it deviations from the average whose root mean square value is $N^{\frac{1}{2}}$. These deviations set a limit to the accuracy with which the average number N may be determined. By the same token they set a limit to the smallest change in N that may be detected. Let this smallest change in N be denoted by ΔN. Thus, ΔN is of the order of $N^{\frac{1}{2}}$. The particular constant of proportionality will depend from probability considerations on the certainty asked for in detecting ΔN. With the above definitions, several relations can be written out of hand.

$$\text{Scene brightness} \equiv B \sim \frac{N}{h^2} \tag{1}$$

$$\text{Threshold contrast} \equiv C = \frac{\Delta B}{B} = \frac{\Delta N}{N} \sim N^{-\frac{1}{2}} \tag{2}$$

$$B \sim \frac{1}{C^2 h^2} \tag{3}$$

or

$$BC^2h^2 = \text{constant} \tag{3a}$$

or

$$BC^2\alpha^2 = \text{constant} \tag{3b}$$

where α is the angle subtended by "h" at the lens.

Eq. (3*b*) is already the characteristic equation of the ideal pickup device. The constant term on the right will be evaluated shortly. It contains in straightforward fashion, the lens parameters, exposure time, and quantum efficiency of the device. The meaning of eq. (3b) is this. Let any two of the variables (B, C, α) be specified. Eq. (3b) then sets the threshold value for the third. For example, if the scene brightness and contrast are specified, eq. (3b) gives the smallest angular size that can be resolved. Conversely, if one knows the angular size and contrast of a test object, eq. (3b) determines the minimum scene brightness required for detecting it. It is to be noted that the resolution (α) or half tone discrimination (C) improve only as the square root of the scene brightness. It is to be noted also that the validity of eq. (3b) rests only on the countability of the absorbed quanta and the random character of the absorption process. The first was the necessary condition for ideal performance; the second is a well established experimental fact. Since no special mechanism was assumed, eq. (3b) can apply equally to the eye, to photographic film, and to the great variety of television pickup tubes already discussed insofar as each of these devices can make an accurate count of the absorbed quanta.

Eq. (3b) is the basis for a performance scale according to which the performance of any particular device, whether it be eye, film, or pickup tube, may be judged. It is a relatively straightforward operation to measure the smallest angle that a given device can resolve at a specified scene brightness and contrast. The product of these three quantities, as in eq. (3b) yields a number whose value locates the position of the given device on the performance scale. This scale is an absolute one in the sense that performance is limited only by the statistical fluctuations in the rate of absorption of light quanta, a fundamental and unavoidable limitation. Whatever happens subsequent to the primary process can only deteriorate, or at best maintain, the performance already computed in terms of the primary process. Subsequent elements *can not improve performance.*

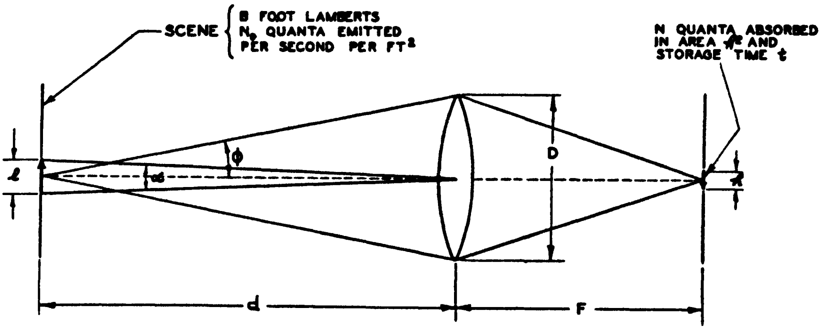

SCENE
B FOOT LAMBERTS
N_0 QUANTA EMITTED PER SECOND PER FT²
N QUANTA ABSORBED IN AREA h^2 AND STORAGE TIME t
l
ϕ
α
D
h
d
F

FIG. 1.—Geometric relations used to compute number of absorbed quanta as a function of number of emitted quanta.

The performance scale, based on eq. (3b), will be more significant if the constant term on the right is separated into its component factors. To do this, reference is made to Fig. 1. Consider an element of side length "l" in the scene. If the scene brightness is such that N_0 quanta are radiated/square foot/second, the total number of quanta radiated by this element per second will be N_0l^2. Of this number, only a fraction will be intercepted by the lens and, for a Lambert distribution, the number passing through the lens per second is:

$$N_0l^2 \sin^2 \phi$$

This number of quanta will also be incident on the element of side length "h" which is the image of the scene element of side length "l." The number of quanta actually absorbed at "h" in the exposure time of the device is

$$N = \theta t N_0 l^2 \sin^2 \phi \tag{4}$$

where θ is the quantum yield or absorption coefficient* and t is the exposure time. From the relations (see Fig. 1):

$$l = \frac{d}{F} h$$

$$\sin \phi \doteq \frac{D}{2d}$$

and

$$\alpha \doteq \frac{h}{F}$$

eq. (4) may be rewritten in the form

$$N = 1.4 N_0 D^2 t \theta \alpha^2 \times 10^{-10} \tag{5}$$

where α is expressed in minutes of arc, and D is expressed in inches. Using the equivalence: 1 lumen of white light = 1.3×10^{16} quanta/second,

$$\frac{N_0}{1.3 \times 10^{16}} = B \text{ foot-lamberts}$$

and, from eq. (5)

$$N = 2 B D^2 t \theta \alpha^2 \times 10^6$$

or

$$B = 5 \frac{N}{D^2 t \theta \alpha^2} \times 10^{-7} \text{ foot-lamberts} \tag{6}$$

From eq. (2), the threshold contrast is related to $N^{-\frac{1}{2}}$ by a constant whose value is yet to be determined. This constant is the threshold signal-to-noise ratio and, as mentioned, is a function of what one calls threshold. That is, if one asks that each observation have a 90% chance of being correct, the threshold signal-to-noise ratio will be higher than if one were satisfied with a 50% chance. From visual observations to be cited later, a reasonable value for the threshold signal-to-noise ratio appears to be greater than unity and probably in the neighborhood of 5. For the present this constant will be introduced by the letter "k." Thus, inserting k^2/C^2 for N in eq. (6) we get the complete form of the characteristic equation,

$$B C^2 \alpha^2 = 5 \frac{k^2}{D^2 t \theta} \times 10^{-7}$$

or, expressing C as a per cent contrast, $\left(C = \frac{\Delta B}{B} \times 100\right)$ we get:

$$B C^2 \alpha^2 = 5 \frac{k^2}{D^2 t \theta} \times 10^{-3} \tag{7}$$

---

* The absorption coefficient is restricted to those quanta that give rise to countable events.

An inspection of the right hand side of eq. (7) shows that the only parameter that differentiates the performance of one ideal pickup device from that of another is the quantum yield (θ) of the primary photoprocess. This statement will probably be clearer if eq. (7) is rewritten in the following form:

$$\frac{5k^2}{BC^2\alpha^2D^2t} \times 10^{-3} = \theta \tag{8}$$

The five parameters needed to specify a given set-up are in the denominator of the left hand side of eq. (8). Their names and units are: scene brightness (B foot-lamberts), contrast of test object $\left(C = \frac{\Delta B}{B} \times 100\%\right)$, angular size of test object (α minutes of arc), lens diameter (D inches) and exposure time (t seconds). When any four of these are arbitrarily specified, the threshold value for the fifth is given by eq. (8). Alternatively, if all five parameters are known, a value for the quantum yield (θ) is thereby determined. If the operation of the device is ideal, that is, if all of the absorbed quanta are counted, the value for θ determined in this way *is* actually the quantum yield of the primary photoprocess. If, on the other hand, the operation departs from ideal operation—and the ways in which such departures may occur are both manifold and frequent—the value for the quantum yield so determined is *less* than the quantum yield of the primary photoprocess. In this event, it may still be looked upon as an index to performance. To recapitulate: a given value for θ computed from eq. (8) may correspond to an ideal device having the same value of θ for its primary photoprocess or to a nonideal device having a larger value of θ for its primary photoprocess, the nonideal characteristic being responsible for the lower *computed* value of θ.

Eq. (8) defines the absolute performance scale to be used later in evaluating the performance of various pickup tubes, photographic film, and the eye. The scale extends from $\theta = 0$ to $\theta = 1$. The value $\theta = 0$ means no transmitted picture. The value $\theta = 1$ means not only ideal operation but absorption of all incident quanta as well. It should be clear at this point that a single value of θ may be used to designate the performance of a device only over a limited domain of the five parameters B, C, α, D and t. In fact, in the usual case, θ is a continuous function of these parameters. No device, for example, can resolve arbitrarily small angles. Nor can any device discriminate arbitrarily small contrasts. Nonstatistical considerations set limits to the smallest values attainable by either of these parameters. Also, the range of times over which a device can store up information has an upper limit set by the mechanics of the device and not by statistical limitations.

It would indeed be a thorough-going paper that explored the performance of various pickup devices for variations of all five of the parameters of eq. (8). Such completeness will not be attempted here. Rather, only one of the parameters, the scene brightness, which is of greatest interest, will be selected for examination. Reasonable fixed values or ranges of values of the other parameters will be taken. Some remarks will be made in passing about the dependence of θ on these other parameters but with no object of making the review exhaustive. The parameter, scene brightness, is of special interest because few devices can maintain a high level of performance over an appreciable range of scene brightnesses. The eye is, perhaps, the outstanding example. Other devices may match or even exceed the eye over narrow ranges of scene brightness but, as yet, they lack the flexibility of the eye to maintain performance both at very high and at very low lights.

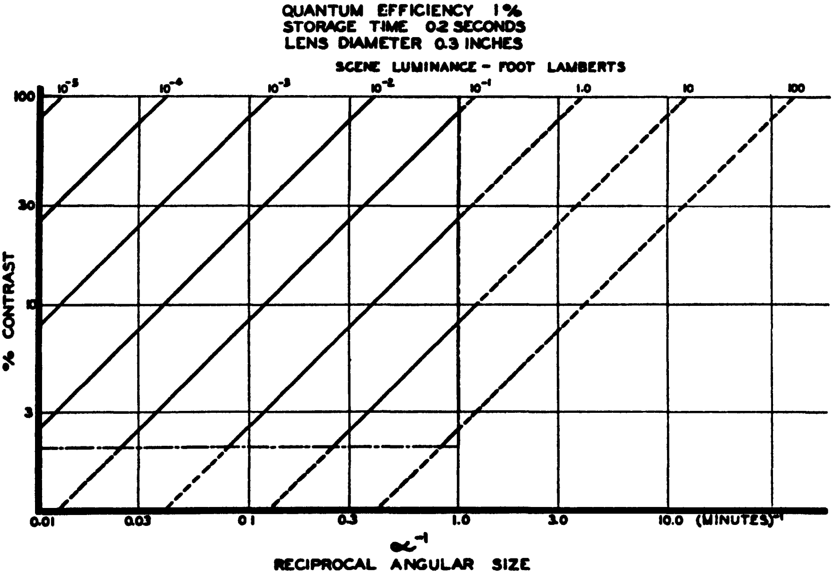


FIG. 2.—Performance curves for an ideal pickup device.

Before estimating the performance of various selected pickup devices, it will be well to try to tie the characteristic eq. (8) closer to reality, both by inserting specific values and by outlining an experimental arrangement which satisfies this equation in simple fashion. This arrangement will also furnish an experimental value for threshold signal-to-noise ratio, k.

In Fig. 2, the threshold contrast has been plotted against the reciprocal angular size for a large range of scene brightnesses. The values for quantum efficiency, storage time, and lens diameter shown in this figure were selected to approximate the human eye over part of its range of operation. The value 5 was assumed for the threshold signal-to-noise ratio (k) in computing these curves. The dash-dot lines indicate the limits to the smallest angular size and contrast that can be detected, which limits are set variously by the particular mechanisms of particular devices and are not set by statistical fluctuation considerations. For this reason, the curves are dotted outside the dash-dot boundaries.

By way of example, Fig. 2 shows that, at a scene brightness of 1 foot-lambert, an ideal device having a quantum yield of 0.01, a storage time of 0.2 seconds and a lens diameter of 0.3 inches can just detect a contrast of 30% for objects subtending 1 minute of arc at its lens. If smaller contrasts, say around 3%, are to be detected, the object must be larger and subtend an angle of about 10 minutes. At very low scene brightnesses, Fig. 2 shows that only a poor quality picture may be transmitted (poor resolution and poor discrimination for half tones) corresponding to the obscured pictures that the eye sees at night. The poor quality of these pictures is a direct consequence of the lack of sufficient light quanta and cannot be avoided by any "improvements" in design other than improved quantum yield in the primary photoprocess.

Lest the present discussion be dominated by a purely speculative tone, an actual physical system will be described in the next section which brings out the essential features of ideal performance.

VI. An Experimental Realization of Ideal Performance

An early and simple means for generating a television picture is embodied in the so called light-spot scanner. Fig. 3 shows the essential parts of this system. The scene to be transmitted is "illuminated" by a small sharply focused spot of light which scans the scene in the customary manner in a series of parallel lines. At each point in the scene some of the scannng spot light is reflected and is picked up by a photocell. The light and shade of the scene are conveyed by the variations in scattered light as the spot scans over its surface, and the variations in scattered light are in turn conveyed by variations in photocurrent in the photocell. A viewing tube or kinescope is connected to the photocell in such a way that, as the original light spot scans its scene, the kinescope beam scans its luminescent screen in synchronism and, by variations in beam current controlled by the output of the photocell, reproduces the original scene. The light-spot scanner is restricted to those scenes that may be conveniently illuminated by a scanning spot of light. Recent developments in phosphors and photomultipliers have revived interest in the light-spot scanner for those applications (Sziklai, Ballard, and Schroeder[1]). Its operation is simple and it provides a picture free from most of the spurious signals frequently encountered in storage type pickup tubes.

The particular virtue of the light-spot scanner for the present discussion is that it offers a ready means for demonstrating the properties of ideal performance. For example, the gain of the electron multiplier in the photocell is sufficiently high that each quantum that is absorbed at its photocathode (and liberates a photoelectron) can be made visible on the kinescope as a discrete speck of light. Thus, at extremely low scene brightnesses, one can actually and easily count the number of "quanta" in the reproduced picture.

The special test pattern which served as subject for the light spot scanner and which measured its performance is shown in Fig. 4. This pattern is a materialization of the plot in Fig. 2. The discs along any row have the same contrast but vary in diameter step-wise by a factor of 2 for each step. The discs in any column have the same diameter but vary in contrast also step-wise by a factor of 2 for each step. At moderate or low illumination not all of the discs in this pattern can be seen or transmitted by a pickup device. The demarcation between those discs that can be seen and those that cannot should be a 45° diagonal if the picture is limited only by noise whose distribution in frequency is uniform (e.g., shot noise in a temperature limited beam of electrons). As the illumination of the test pattern is varied the line of demarcation should move from one diagonal of discs to the next for a factor of 4 change in illumination.

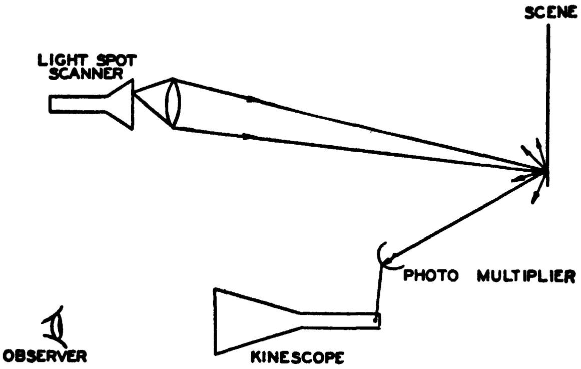


FIG. 3.—Light-spot scanner arrangement.

The series of photographs in Fig. 5, is a series of timed exposures of the kinescope taken as the light-spot scanner scanned the test pattern at the usual television rate of 30 pictures/second. For technical convenience the exposure time of the camera was varied rather than the scene brightness since, according to eq. (8), it is only the product Bt that is significant.

The exposure times in the first few pictures are low enough to see the separate "quanta" (or their corresponding specks of light on the kinescope). As the exposure time is increased, more and more of the test pattern becomes visible. In particular, for a range of over 200 in exposure times, the demarcation shifted one diagonal of discs for each factor of 4 increase in time. Also, the demarcation on any one picture is approximately along a 45° diagonal. These two observations confirm the significant features of eq. (8), namely that at any one scene brightness, threshold contrast is proportional to the reciprocal angular size and that both the reciprocal contrast and the reciprocal angular size vary as the square root of the scene brightness.

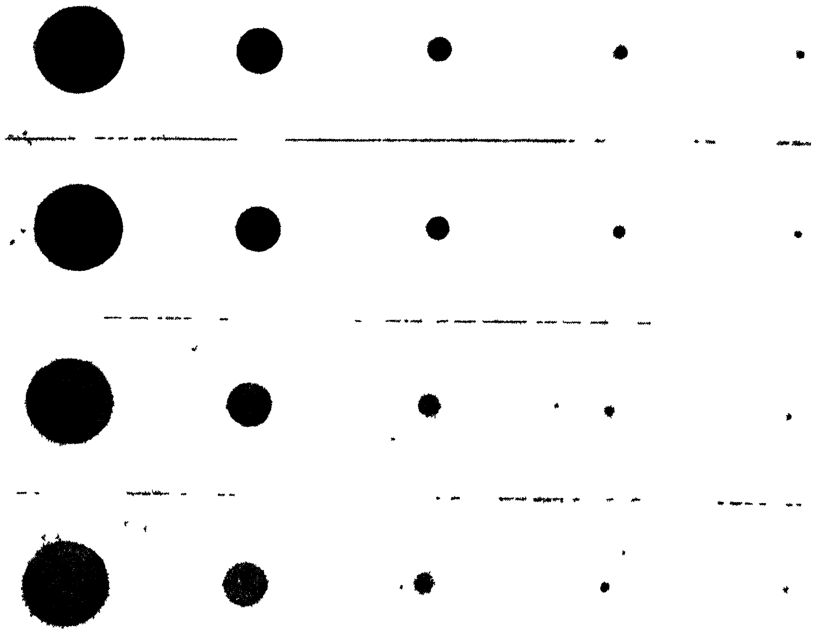


FIG. 4 —Test pattern for measuring performance.

The series of photographs in Fig. 5 provide also a means for estimating the exposure time of the eye and the value of k, the threshold signal-to-noise ratio.

Concerning the exposure time of the eye, it was found that, of the series of photographs of varying exposure times, the picture that most nearly matched the direct visual impression of the kinescope was that for

0.2 seconds. This test was made actually more quantitative by placing a mask over the kinescope and counting the average number of specks visible in a small area. The particular photograph was then selected that had the same average number of specks in the same area. The exposure time of the eye determined in this manner is in satisfactory agreement with well known and more thorough determinations already published.

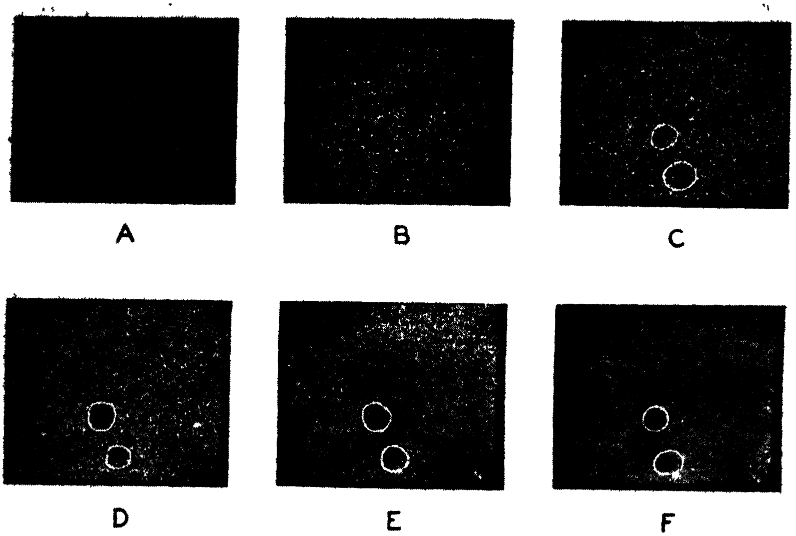


Fig. 5.—Pictures transmitted by the light-spot scanner in Fig 3 of the test pattern in Fig. 4. The pictures were photographed from the kinescope using camera exposure times starting from Fig. 5A, of $\frac{1}{16}$, $\frac{1}{4}$, 1, 4, 16, and 64 seconds respectively. The ultraviolet component of the scanner light was used in order to get a short luminescent decay time. The circled areas are blemishes that were not apparent under visible light and have no connection with the test.

Concerning the threshold signal-to-noise ratio, "reasonable" assumptions, for lack of experimental evidence, have been made (Rose,[2] DeVries[3]) that its value is unity. A recent direct determination (Schade[4]), however, sets the value of k in the range of 3–6, depending on the viewing conditions. The present paper has used the value 5. Whatever uncertainty there may be in estimating a value for the threshold signal-to-noise ratio from the photographs in Fig. 5, that value is definitely greater than unity and is in the range of 3–6. This is surprisingly high since mathematical analysis suggests a value close to unity. One way of reconciling the discrepancy is to assume that the eye in looking at the pictures in Fig. 5 does not make full use of the information presented. That is, the eye (and brain) may count only a fraction of the "quanta" present. If one used a machine to do the counting and arranged to ring a bell whenever the count warranted, one might approach a threshold ratio of unity. Since, however, one usually chooses to do his own counting by the normal visual process, the experimentally determined value of 5 for threshold signal-to-noise ratio is significant.

The operation for computing a threshold signal-to-noise ratio from Fig. 5 is as follows. Select the smallest black (not grey) disc that is visible in any one of the photographs in which the specks are countable. Transpose the outline of the black disc to the surrounding uniform white area. Count the average number of specks in the outlined area. This average number is the signal; the square root of the average number is the root mean square noise. The threshold signal-to-noise ratio is then also the square root of the average number. A similar operation can be carried out to get the same threshold signal-to-noise ratio using the outline of a threshold grey dot. In this instance the signal is the average number multiplied by the contrast (not per cent contrast) and the noise is the square root of the average number. The results of these operations, as already mentioned, lead to a value of k in the neighborhood of 5.

The light-spot scanner was used also to make a direct comparison estimate of the performance of the eye. Even though the light-spot scanner is a nonstorage pickup device and has the characteristic insensitivity of such a device, it turns out that the average illumination on the scene as provided by the scanner is as low as it would be for an ideal storage type pickup tube operating with uniform continuous illumination. The reason is that the scanner illuminates the scene only when and where that illumination can be used. There is no waste illumination. Hence, if one measures the average scene brightness for a scene illuminated by a light-spot scanner and observes the signal-to-noise ratio of the transmitted picture, these quantities correspond exactly to the performance of an ideal pickup device having the same quantum efficiency as the photosurface of the photomultiplier. To complete the parallel, the diameter of the photocathode of the multiplier replaces the diameter of the lens in eq. (8). The exposure time of the system is the exposure time of the observer (human or instrumental) that looks at the final picture on the kinescope. The direct comparison between photomultiplier and eye was further facilitated by the fact that the useful area of the multiplier photocathode is about equal to the diameter of the dark-adapted eye and that the spectral distribution of the scanner light overlapped about equally the response curves for eye and multiplier. The test was carried out by having the photomultiplier and human eye both look at the original scene at the same viewing distance. The human observer then looked at the kinescope to compare what the photomultiplier saw with what the human eye could see directly on the original scene. Using the test pattern of Fig. 4 it was interesting to find the eye and photomultiplier performance very closely the same in the test range of 10^{-2} to 10^{-4} foot-lamberts. Since the picture transmitted by the photomultiplier was obviously limited by fluctuation noise in its primary photoprocess and since the quantum efficiency of the photomultiplier surface was about 2%, the quantum efficiency of the eye must also have been 2% if it were limited by fluctuation noise in its primary photoprocess or greater than 2% if some less fundamental limitation were present.

VII. Performance of Selected Pickup Devices

The object in the following sections will be to trace the performance of various representative pickup devices as a function of scene brightness, using the absolute scale of performance already discussed. Because the range of scene brightnesses to be considered is over ten orders of magnitude and the range of performance values (θ) over seven orders of magnitude, the precision attainable in a single composite plot will not be large. What is aimed at in this survey and the corresponding plot is a picture of the areas of performance not yet covered by any device as well as the areas already covered by the eye but not yet by any man-made devices. In brief, it is a picture of performance possibilities yet to be realized.

1. *Human Eye*

To locate the performance of the eye, the six parameters on the left hand side of eq. (8) need to be known. There are data in the literature (Cobb and Moss[5]; Connor and Ganoung[6]; Blackwell[7]) relating scene brightness (B), threshold contrast (C) and angular size (α) for the eye. These data cover the following ranges: B, 10^{-6}–10^{2} foot-lamberts, C, 1–100% contrast, and α, 1–100 minutes of arc. In addition, less complete observations have been made above 10^{2} foot-lamberts and below 10^{-6} foot-lamberts. For example, above 10^{4} foot-lamberts pain sets in and below 10^{-7} foot-lamberts vision ceases completely. There are also data for the pupil diameter (D) as a function of scene brightness (Reeves[8]). If the storage time (t) of the eye is taken to be 0.2 seconds, and independent of scene brightness, and if, for the threshold signal-to-noise (k), the previously discussed experimental value of 5 is taken, the roster of information necessary to compute performance (θ) is complete. One further qualification needs to be made. It is found from Blackwell's data that at any fixed value of scene brightness, the threshold contrast is only approximately proportional to the reciprocal angular size as demanded by eq. (8). One can, however, for the purpose of this plot, select the best performance at each value of scene brightness (see also Rose[9]). This procedure was used to compute the curve of θ vs. B for the eye in Fig. 6.

The most striking feature of this curve is the exceedingly large range of scene brightnesses, namely 10^{-6} to 10^{2} foot-lamberts, over which the eye maintains a high level of performance. It is striking because no single man-made device has succeeded in covering this range. Even more, the ensemble of man-made devices falls far short of covering it. The experience of those who have tried to build the same flexibility and excellence of performance into electronic devices can only suggest the artfulness of the mechanisms that in some way have been engineered into or have selectively survived in the eye. Two of these mechanisms in particular deserve mention.

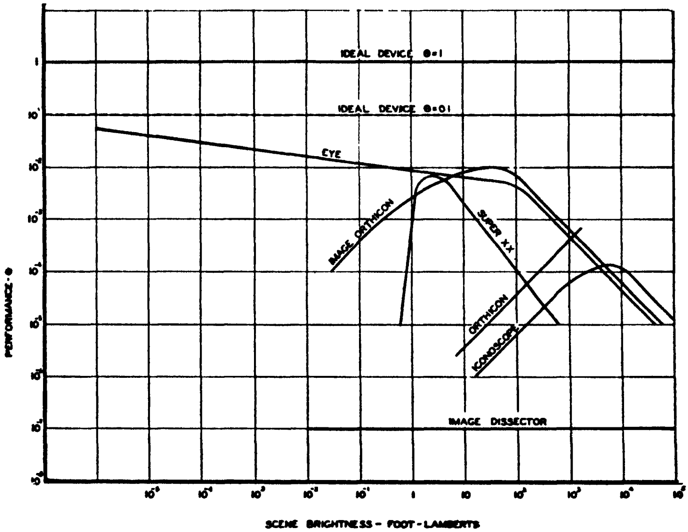


FIG. 6.—Performance curves for selected pickup devices.

There is good evidence that the eye requires more than a single absorbed quantum to generate a visual sensation. If, now, the rods and cones had no interconnections but acted completely independently in their reception and transmission of information to the brain, the eye would rapidly fail to "see" when the concentration of absorbed quanta fell below the concentration of rods and cones. This, in fact, is just the reason that photographic film cuts off sharply towards low scene brightnesses. The eye, then, must have a mechanism that can pool the information from the separate rods and cones when such pooling is advantageous for seeing and can separate the operation of these rods and cones when, as at high lights, there is an abundance light quanta. In brief, the mechanism must automatically control the interconnections of the retinal elements to get the maximum information out of the incoming stream of quanta.

The second interesting mechanism has to do with the existence of a variable-gain element located between the retina and the nerve fibers that carry pulses to the brain. A gain element of some sort is needed in order to raise the energy level of the few quanta required for a threshold sensation up to the much higher energy level of the subsequent nerve pulse. If the gain of the element were constant, however, the nerve fibers would be called upon to pass visual sensations varying (according to Fig. 6) by about eight orders of magnitude. While such a design is not inconceivable, it is not especially elegant. By way of example, if one wants a device capable of recording the variation of a quantity in ten thousand discrete steps, it is simpler to compose the device of two elements each capable of one hundred different values, such that their product yields the ten thousand discrete steps than to try to design the ten thousand steps into a single element. Whatever the validity of this argument, there is good reason to believe from observations on dark adaptation, that the gain of the above mentioned element varies automatically with scene brightness (Rose[9]). The delay in "resetting" during the transition from high lights to low lights would correspond to the time taken for dark adaptation. Also, fluctuations (noise) in the primary photoprocess (absorption of light quanta at the retina) should be visible if the computed value of θ (eq. 8) is equal to the value of θ observed or computed directly for the retina. Hecht,[10] for example, by a statistical analysis of visual observations near threshold, arrives at a value of θ of about 0.1. Hecht's value for blue light should be divided by a factor of 3 for comparison with the white light observations of Fig. 6. The agreement is close enough to suggest that the variable gain is automatically set to make fluctuations just visible. From necessarily subjective observations the writer confirms this conclusion at least at low lights around 10^{-4} foot-lamberts.

This discussion leaves out the mechanisms, probably still more ingenious, that are responsible for color vision.

2. Photographic Film

The millions of years spent in evolving the human eye have yielded a device with a high level of performance over an enormous range of light intensities. The one hundred or more years spent in the development of photographic film have resulted in the same level of performance but confined to a very narrow range of light intensities. With adequate light, pictures produced photographically are remarkable for their uniformity, resolution, rendition of tonal values, and freedom from distortions and spurious effects. As the scene brightness is reduced, however, a relatively sharp threshold is reached below which no photograph, not even a poor one, is obtained. This abrupt departure from ideal performance results mainly from the fact that more than one absorbed quantum is needed to make a photographic grain developable. When the density of absorbed quanta falls below the density of grains, the probability that more than one quantum will be absorbed by a grain rapidly approaches zero. If film could pool its grains, as the eye does its retinal elements, it also could continue to record pictures and maintain its performance level at low lights.

A given film departs from ideal performance, but more slowly, also towards high lights. This departure becomes clear if one first exposes a film so that about half its grains are rendered developable, that is, to a latent density of about 0.3. Now, if the film is exposed further, at least half of the new exposure is wasted on grains that have already been made developable. As the exposure continues, the waste increases and so does the departure from ideal performance. The finite number of grains, and correspondingly finite signal-to-noise ratio, further restricts the approach to ideal performance at high lights.

The relatively sharp threshold towards low lights and the progressive inefficiency of exposure at high lights leads to an optimum of performance for a given film in the neighborhood of a density of 0.3. This optimum can be shifted along the scene brightness scale (Fig. 6) by choice of films with different grain sizes. A given film, however, for fixed values of lens speed and exposure time, has a rather narrow working range.

To locate the curve for Super XX film shown in Fig. 6, data from Jones and Higgins[11] giving the measured signal-to-noise ratios as a function of area on the film and film density were used. Also, a lens diameter of 0.3 inches and an exposure time of 0.2 seconds were selected in order that the depth of focus and speed of operation be comparable with the same properties of the eye. A different choice of lens diameter or exposure time would not change either the shape of the curve or its vertical position on the performance scale. It would merely shift the curve rigidly along the brightness axis. Because the location of the point of departure from ideal performance at the low light end depends upon the relative density of light quanta and grains at the film, it is necessary also to specify the focal length of the lens. This insures a one to one correspondence between scene brightness and illumination at the film. The focal length is chosen to give a commonly used angle of view of 25°. Thus, for 35-mm. film, the vertical height of whose picture is about 16 mm., the focal length would be 1.5 inches. Another choice of focal length would, like variations in lens diameter or exposure time, only shift the curve rigidly along the brightness axis.

3. Television Pickup Tubes

The following tubes have been selected from amongst a great variety of possible tubes, first because they have all been realized and put into service, and second because historically and logically they represent successive steps towards attaining ideal performance. A more detailed discussion of the mechanics of their operation can be found in the cited references, or in Zworykin and Morton,[12] or in a recent summary by Zworykin and Ramberg.[13]

To compute the pickup tube curves, a lens diameter of 0.3 inches and an exposure time of 0.2 seconds were uniformly chosen for ready comparison with the eye. Also, the focal length of the lens was selected to give an angle of view of 25°.

It is interesting to note that the choice of 0.2 seconds for the exposure time is not inconsistent with the usual television frame time of 0.03 seconds. This longer exposure time is set by the final observer at the kinescope. The human eye, for example, integrates about six successive television frames in its storage time of 0.2 seconds. The same holds for the viewing of motion picture film. The result is an improved signal-to-noise ratio over what one would have gotten if he actually observed separate frames.

It is well to point out also that for pickup tubes the focal length of the lens may in general be varied with no effect on the shape or position of their performance curves, providing the linear size of target is varied in the same proportion. This differs from film for which the intensity of illumination rather than the total illumination on the target is critical.

a. Image Dissector. The image dissector (Fig. 7) (Farnsworth,[14] Larson and Gardner[15]) rapidly attained the theoretical limit of performance for nonstorage type pickup tubes. Since, however, the lack of storage already puts this limit about five orders of magnitude (number of separate picture elements) below that for storage type tubes, the image dissector is confined to applications where sensitivity is not important. The transmission of motion picture film or slides is one such application.

The operation of the image dissector, except for its lack of storage, is an excellently simple example of the counting process that is the ultimate basis of all picture pickup devices. The picture to be transmitted is focused on a conducting photocathode. The photoelectrons are then focused at 1:1 magnification at the other end of the tube. Here the tube is closed except for a small aperture, the size of the desired picture element. The entire picture is deflected across the aperture so that the aperture can explore each portion of it every thirtieth of a second. The electrons that pass through the aperture are multiplied by an electron multiplier whose gain can be high enough to count individual electrons. At any one moment, the electrons from only one picture element pass through the aperture. If one counts the total number of electrons that pass through the aperture in the time it takes to scan a picture element, that number is a measure of the signal and its square root is actually the signal-to-noise ratio referred to a picture element. It is also the signal-to-noise ratio that would be measured in the conventional manner by observing the current and noise power with appropriate meters.

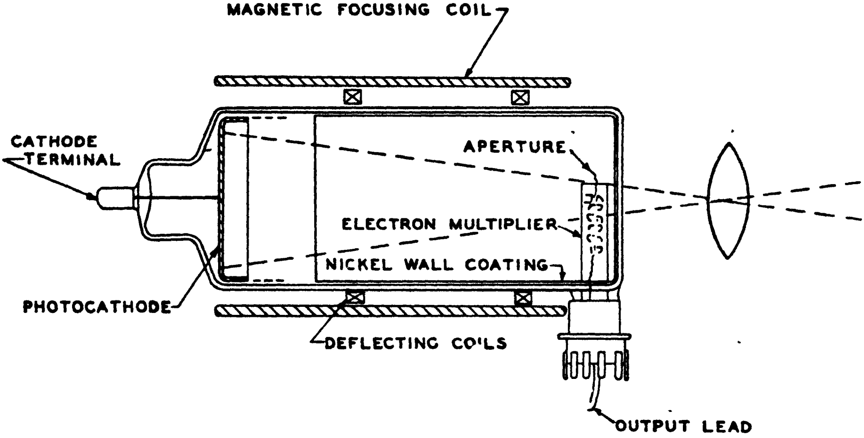


FIG. 7.—Image dissector.

In addition to the five orders of magnitude loss in performance due to lack of storage, the image dissector is down another two orders of magnitude (as are all the other pickup tubes) because its photocathode has a quantum yield of about 0.01. The resulting curve in Fig. 6 shows a constant level of performance over its full light range, but seven orders of magnitude down from the theoretical maximum.

b. Iconoscope. The iconoscope (Fig. 8) (Zworykin, Morton, and Flory[16]) made the important step of introducing storage into pickup tubes. Although the iconoscope does not make full use of storage and is somewhat handicapped by a characteristic nonuniform shading pattern, it made possible the transmission of high quality "live" pictures. In fact, one of the very elements in its operation that is responsible for its inefficiency also leads to a reproduction of tonal values that has been difficult for other more sensitive tubes to match. This element is the incomplete collection of photoemission at high lights.

The optical image to be transmitted is focused on a photosensitive and insulating target. A charge pattern is stored which, when scanned by a constant current beam of electrons, modulates the escape of the secondary electrons. An amplifier is connected to the signal plate on the backside of the target and observes the variations in escape of secondary electron current. Equilibrium potentials are maintained by a rather involved process of redistribution of some of the secondary electrons over the target. While the saturated photosensitivity of the target may be about 10 microamperes/lumen, the operating photosensitivity is much less owing to lack of saturation. The operating photosensitivity is only 1 or 2 microamperes/lumen at low lights and a third to a tenth of that at high lights. Further, the noise associated with the picture is not the noise in the escaping stream of secondary electrons but the much larger fixed noise characteristic of the amplifier. Thus, low-level signals (or low counts) are obscured proportionately more than high-level signals. For this reason, the performance curve (Fig. 6) rises towards the high-light end.

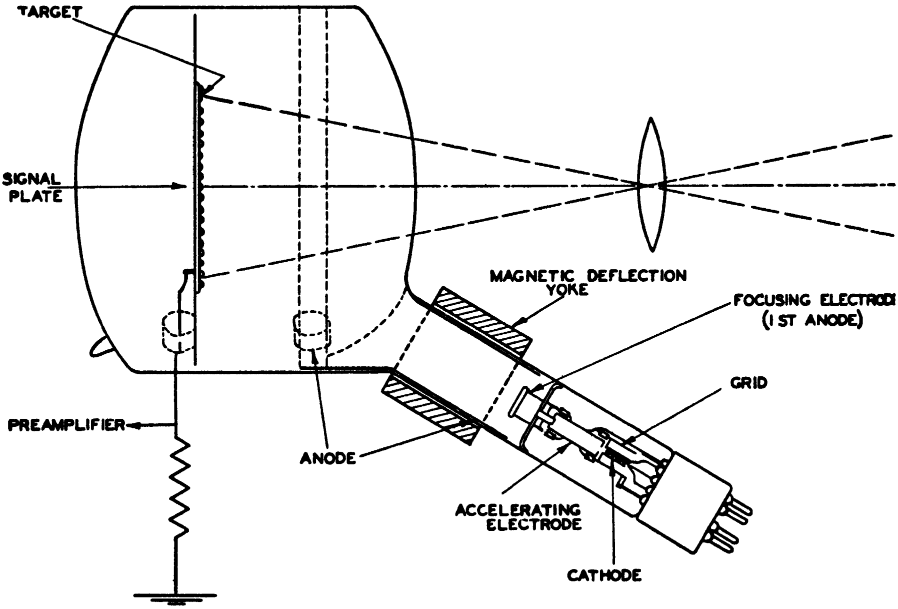


Fig. 8.—Iconoscope.

To compute the performance curve for the iconoscope, its known signal- vs. light curve was used together with the noise current value (2×10^{-9} amperes) for a television amplifier with a pass band of 5 Mc. A more accurate appraisal would take into account the fact that the noise is not uniformly distributed over the pass band but is peaked at the high frequencies. The effective noise current may accordingly be as much as a factor of 3 lower (Schade[4]).

The combination of a 0.3-inch lens diameter and a focal length of 9 inches (for 25° angle of view) leads to a lens speed of $f/30$ and also to the location of the iconoscope curve at exorbitantly high values of scene brightness. This is legitimate for the present purpose of showing what scene brightness the iconoscope would need if it were to match the depth of focus of the eye. In practice, however, the lens diameter is opened up at the expense of depth of focus in order to bring the operating curve down to reasonable values of scene brightness.

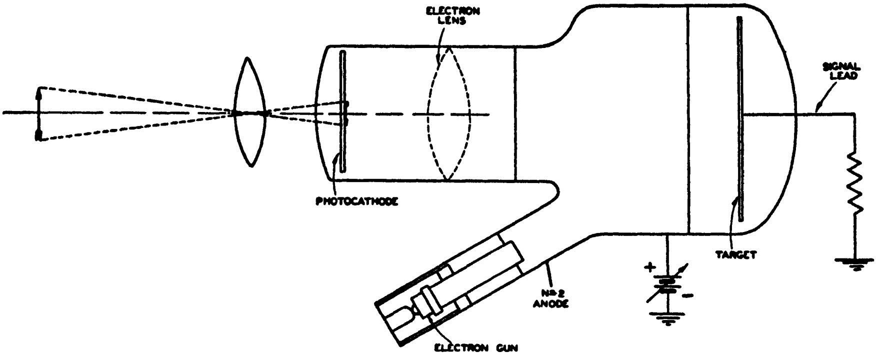


FIG. 9.—Image iconoscope.

c. Image Iconoscope. The image iconoscope (Fig. 9) (Iams, Morton, and Zworykin[17]) differs from the iconoscope by having an electron image rather than an optical image focused on the target. The electron image originates from a conducting photocathode on which the optical image is focused. Sensitivity gains are realized because the conducting photocathode can be made more sensitive than the insulating target surface and because the electron image is amplified by secondary emission at the target. The resultant sensitivity is five to ten times that of the iconoscope. For the rest, its operation is substantially that of the iconoscope. Its performance curve lies in the immediate neighborhood of that for the orthicon, which is about to be described. Its curve is not included in Fig. 6.

As mentioned earlier, the same limit of pickup tube sensitivity may be reached by multiplication of the electron image as by multiplication of the signal current delivered by the scanning beam. It has been technically more convenient, however, to get large gains by the latter process.

d. Orthicon. The orthicon (Fig. 10) (Rose and Iams[18]) avoided two prominent limitations of the iconoscope and at the same time introduced problems peculiar to its own design. The spurious shading pattern and the incomplete utilization of storage of the iconoscope both result from the complex redistribution of secondary electrons generated by the high velocity scanning beam. It was well recognized that a low velocity scanning beam would avoid these defects. Electrons from the scanning beam would then be deposited only where a positive picture charge pattern existed and in an amount equal to the positive charge. There would be little or no interchange of electrons between parts of the target. Also, a strong collecting field could be set up to saturate the photo emission from the target.

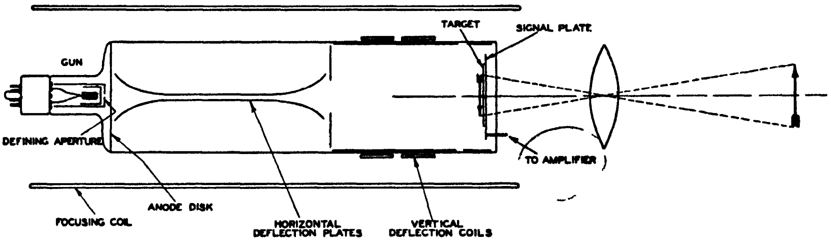


FIG. 10.—Orthicon.

While the virtues of a scanning beam of low velocity electrons were known, it was also appreciated that such a beam was difficult to control. The beam could easily be defocused or deflected by stray charges in the tube and even by the picture charge on the target. Further, the beam had to be deflected in such a way that its angle of approach to the target was at least uniform, and preferably perpendicular, over all parts of the target. The orthicon, by its use of a uniform magnetic focusing field extending the full length of the tube, presented one useful solution to these problems. In the earlier tubes, the slow speed vertical deflection was accomplished by a pair of deflection coils while the high speed horizontal deflection made use of a pair of specially shaped plates. Improvements in deflection circuits made possible a later and simpler design in which both vertical and horizontal deflections were carried out with deflection coils.

As in the iconoscope, the optical image is focused on a photosensitive and insulating target surface. A positive charge pattern is thereby built up and stored. The scanning beam approaches the target at near zero velocity. Where no positive charge is present (dark parts of the picture) the beam reverses direction without striking the target. Where positive charge is present, some of the electrons in the beam land on the target and in an amount just equal to the positive charge. The semitransparent signal plate on the opposite side of the target records the fraction of the beam current that lands and passes this information on to a television amplifier.

The signal out of the tube is proportional to the incident light intensity. The curve extends from low signals whose threshold character is determined by the noise of the television amplifier up to a signal value more or less well defined by the need for maintaining picture quality. That is, at very high lights the scanning beam tends to be defocused and deflected by potential differences on the target.

The transmitted picture is free from spurious shading patterns. This fact, in combination with the linearity of the signal vs. light curve, has recommended the tube for picking up low-light, low-contrast scenes. On the other hand, difficulty has been encountered in trying to squeeze scenes with a wide range of contrasts into the limited signal range of the tube. Another operating problem results from the fact that the cathode potential of the target is a metastable potential. A sudden bright flash of light can charge the target up to the point where it is locked by the scanning beam at the relatively stable level of anode potential. A finite and objectionable amount of time is required to return the target to cathode potential.

To plot the orthicon curve in Fig. 6, a target photosensitivity of 5 microamperes/lumen was assumed together with a focal length of 4.5 inches for the 0.3-inch diameter lens. The vertical height of the target is about two inches. The same television amplifier noise current was used as for the iconoscope to compute signal-to-noise ratios.

Mention should be made here of experimental orthicons that have been designed using electron multiplication of the return part of the scanning beam current. Higher sensitivities and signal-to-noise ratios have been obtained at some sacrifice of target stability at high lights.

e. Image Orthicon. The image orthicon (Fig. 11) (Rose, Weimer, and Law[19]) has incorporated many of the useful operating characteristics of the image dissector, iconoscope, image iconoscope, and orthicon into one tube. It has the freedom from amplifier noise of the dissector, the highlight stability of the iconoscope, the sensitivity increment of the image iconoscope, and the linearity of response and substantial avoidance of spurious shading at low lights of the orthicon. The area under its performance curve* (Fig. 6) is considerably larger than that under any of the other pickup tubes or under Super XX film but still considerably smaller than the area under the eye curve. Over a small range of its performance curve, its θ value from eq. (8) is closely equal to the θ value for its photocathode and also to that for the eye. In other words, it successfully counts all absorbed quanta and achieves ideal operation. Towards low lights the counting process becomes obscured by proportionately more and more spurious counts from unused electrons in the scanning beam and the curve drops away from ideal performance. Towards high lights its signal remains constant with increasing light and again its performance curve departs from ideal performance. Another nonideal characteristic of the tube, and one that cannot be conveniently shown on Fig. 6, is that only the high-light portions of a given picture are properly represented by this curve. The low-light portions of the picture fall along a line of about a 45° slope whose upper end is fixed by the curve in Fig. 6. This is another way of stating that the absolute noise content of the low-light parts of a picture is the same as that for the high lights. For ideal operation the absolute noise content of each part of a picture should be a function of the brightness of that part and in particular should decrease with decreasing brightness.

---

* For effective operation at low lights the target capacitance of the image orthicon should be reduced from its optimum value for high lights.

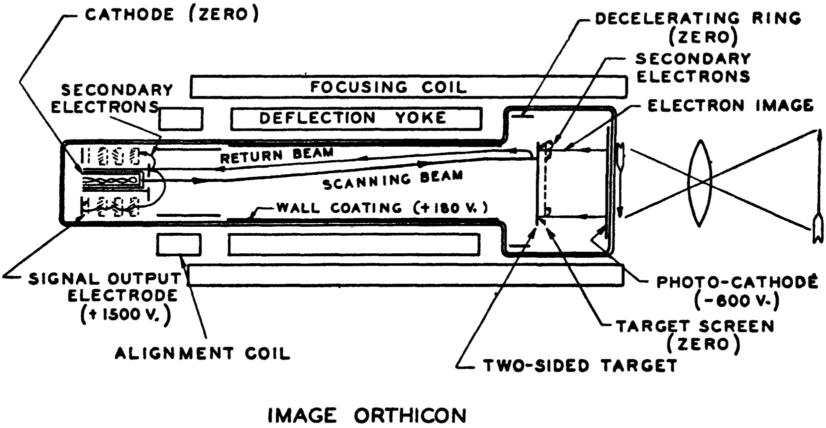


FIG. 11.—Image orthicon.

A rapid trip through the tube will perhaps clarify some of the above points. The optical image is focused on a semitransparent conducting photocathode whose sensitivity, in general, can be made higher than that of an insulating target. The resulting photoelectrons are focused by a uniform magnetic field and accelerated to about 300 volts at which velocity they strike the back side of a two-sided target. Here, by virtue of a secondary emission ratio greater than unity, an amplified positive charge pattern is formed on the target. The secondary electrons are collected by a fine mesh screen located very close to the target. The potential of this collector screen also limits the potential to which the target can be charged by the picture. Stability at high lights is thereby achieved as well as the leveling off of the signal vs. light curve. The two-sided target is a thin sheet of glass. It is thin enough so that the scanning beam on the other side of the target can clearly see the picture charge and deposit, as in the orthicon, an equal amount of negative charge. The resistivity of the glass is adjusted so that these two charge patterns of opposite sign continuously unite and neutralize each other by conduction. The part of the scanning beam that is not deposited on the target returns to an electron multiplier located at and around the electron gun. The gain of the multiplier is high enough to raise the signal and noise level of the electron beam above the noise level of the amplifier to which the tube is connected. While the actual gain of the multiplier may be about five hundred, the useful gain (useful for sensitivity) varies from about twenty for high-light pictures to several hundred for low-light pictures.

From this description it can be seen that when all the photoelectrons are stored on the target (as positive charges by secondary emission) and when most of the beam electrons also land on the target to neutralize the photo charge, the tube can make a fairly accurate count of the number of absorbed quanta. Under these conditions its operation and performance are ideal. At very low lights, however, only a small fraction of the beam electrons, for electron optical reasons, are useful for counting charges on the target. The rest contribute a background noise that obscures the desired count. At high lights, not all the photoelectrons are stored on the target. They could be if the potential of the fine mesh screen collector were made arbitrarily high. But to preserve picture quality, this potential is kept at a reasonably small value limiting the amount of charge that can be stored.

To compute the image orthicon curve in Fig. 6, a photocathode sensitivity of 10 microamperes/lumen was taken and a focal length of 3 inches. Lens diameter and exposure time, as for the other devices, were set at 0.3 inches and 0.2 seconds respectively.

4. Discussion of Performance Curves

The method of plotting and the curves in Fig. 6 are not the most convenient for deciding what pickup device should be used for specified applications. Such would be the case if the performance curves for all of the devices were ideal and if there were no questions of ranges of contrast and angular size to be reproduced, spurious signals, freedom from distortion, stability, size, and so on. The departures from ideal performance, however, and the need for satisfying many side conditions make the choice of a device for a given purpose a matter for careful compromise.

Fig. 6 *was* designed to emphasize these points:

(1) The gap between the performance of actual pickup devices and maximum theoretical performance.

(2) The relatively large range and high level of performance of the eye.

(3) The relatively narrow ranges and generally low level of performance of pickup devices other than the eye.

(4) The inadequacy of defining the sensitivity of *nonideal* devices by a single number. Such a number has meaning only for a specified scene brightness or at most a small range of brightnesses. In general, other qualifications also need to be stated.

(5) The simplicity of the sensitivity scale for *ideal* devices. A single number, the quantum efficiency of the primary photoprocess, suffices (see dotted line marked $\theta = 0.1$).

The items just listed are easily assessed by inspection of Fig. 6. Fig. 6 does not, however, suggest the steady improvement of picture quality that accompanies increased scene brightness. If Fig. 6 were rotated through 45° so that the line for $\theta = 1$ sloped upwards, the desired effect would be obtained. The same would result if $B\theta$ instead of θ, alone, were used to measure performance. In Fig. 12, the curves of Fig. 6 are replotted using $B\theta$ for the vertical axis. On this figure, the various levels of performance, $\theta = 1$, $\theta = 10^{-1}$, $\theta = 10^{-2}$, etc. would be marked out by lines parallel to the line marked $\theta = 1$ but shifted one order of magnitude to the right each time. Also on this figure, and this is its purpose, a horizontal line marks out a line of equal picture quality. By way of example, a horizontal line drawn at $B\theta = 10^{-2}$ intersects all of the curves at different scene brightnesses. Its meaning may be stated as follows: The picture that the eye sees at 1 foot-lambert could be seen also by an ideal device with $\theta = 1$ at 10^{-2} foot lamberts; by an image orthicon or by Super XX at 2 foot-lamberts, by an orthicon at 200 foot-lamberts, by an iconoscope at 600 foot-lamberts and by an image dissector at 10^5 foot-lamberts, all of these devices having the same exposure time and depth of focus as the eye. It is seen here immediately that if one had chosen to draw the horizontal line at another value of $B\theta$, the relative sensitivities of the various devices, owing to their departures from ideal performance, would not have been the same.

Consider another horizontal line drawn tangent to the highest point on the Super XX curve. This also intersects the eye curve at about 3 foot-lamberts and says that the picture quality seen by the eye at 3 foot-lamberts matches the best quality that Super XX can transmit. By quality here is meant half tone discrimination or signal-to-noise ratio. If the eye, then, looks at motion pictures taken with 35-mm. Super XX and if the picture subtends an angle of 25° at the eye, the eye will not be critical of the shortcomings of the film, providing the motion picture screen brightness is 3 foot-lamberts or less. If the screen brightness is raised, the discrimination of the eye exceeds that of the film and the eye becomes conscious of the noise or graininess of the film. Since motion picture screen brightnesses are nearer 10 than 3 foot-lamberts, finer grained films requiring higher scene brightnesses than Super XX are used for good quality pictures.

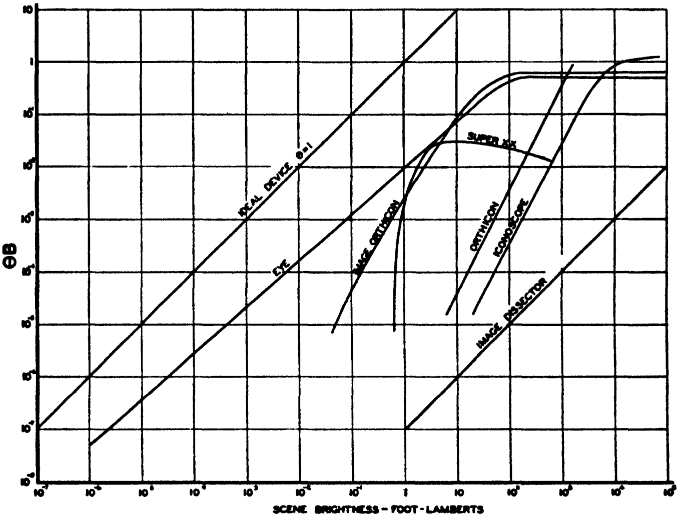


FIG. 12.—Performance curves of Fig. 6 replotted with $B\theta$ as ordinate.

The same considerations hold in somewhat exaggerated form when, for example, the eye looks at kinescope pictures whose brightness may be around 50 foot-lamberts and the pictures are those transmitted by an image orthicon picking up a scene whose brightness may be only a few foot-lamberts. If the television camera were to match the depth of focus of the eye, the resultant kinescope pictures would be exceedingly noisy. The eye would in a sense be looking at its own foot-lambert-quality pictures with a sense of discrimination corresponding to tens of foot-lamberts. If the eye could accomplish this feat more directly it would indeed be critical of its own low light performance. Presumably, the brightness or gain controls in the eye are more nicely matched to the eye's performance than are the arbitrarily variable knobs on a television receiver.

The conditions outlined above—namely, observer looking at a bright kinescope and camera looking at a dim scene—can still result in noise-free pictures if the camera lens is opened up at the expense of depth of focus. This is generally a quite acceptable compromise. The alternative solution, a more sensitive camera, hangs chiefly on the development of more sensitive photocathodes. And all too little is known of the workings of present photosurfaces, let alone improved ones.

The problems of noise visibility just discussed arise frequently and in varied form. They can lead to particularly complex analyses when many parts of a system must be considered. A method of analysis that has wide applicability and that avoids these complications will be outlined in a following section. Another sensitivity problem will be discussed here first.

In a recent paper (Rose[19]) the writer evaluated a figure of merit designed to give the relative sensitivities of photographic film and the human eye. The ratio was about 200:1 in favor of the eye and was based essentially on the facts that the scene brightness and lens area for the motion picture camera that records pictures are each about ten times the screen brightness and lens area for the observer who views them. This ratio (200) would appear to be inconsistent with the curves in Fig. 12 which show film approaching closely the performance of the eye at least near the low-light end of the film curve. A major part of the discrepancy, however, appears to be removable if due account is taken of the relatively sharp threshold for film. For example, if an observer and a motion picture camera both look at a scene whose high lights lie at the point where the curves for eye and film in Fig. 12 are closely matched, these high lights will be reproduced approximately with equal quality. But the low lights of the scene will appear black on the film by virtue of its sharp cut off, while for the eye they will appear with many shades of grey. In fact, the eye very seldom sees true blacks in a scene. If the range from high lights to low lights in the scene is of the order of a hundred fold, the camera would need the same order of increase of scene brightness or exposure in order to avoid rendering black what the eye normally sees as grey. There are other considerations, such as the gamma of the final photographic print, that influence the relative amounts of light required in taking and in viewing motion pictures, but the above orgument would appear to offer the largest factor.

VIII. A CRITERION FOR NOISE VISIBILITY

Noise in a television picture is common enough to be familiar to almost any one who has looked at television pictures. The counterpart of noise in a television picture is graininess in photographic film. But, to the casual observer of motion pictures, graininess is neither frequent nor prominent. That is mainly because film is viewed under conditions more carefully controlled than they are for television pictures. Increase the brightness of the motion picture screen, or let the observer come arbitrarily close to the screen, or use the "fastest" film material and motion picture graininess would be as common as television noise. Whether or not the usual observer is aware of graininess in the pictures he witnesses, is not of primary importance in this discussion. What is to the point is that the existence of graininess has already imposed its restrictions on the brightness of motion picture screens, the minimum approach to such screens, and the scenes that one chooses to photograph. Again, it would be even more difficult to find observers who agreed that they were annoyed by fluctuations in their own visual processes. Yet there is good reason to believe that such fluctuations do limit what can be seen. In brief, the performances of the several devices—television pickup tubes, photographic film, and the eye—are circumscribed by noise, more or less prominently displayed. When two or more of these devices are combined into a system, it is desirable to be able to assess the relative noise contributions of each.

But, in a system involving a scene of variable brightness, a camera of specified performance and adjustable optical geometry, a receiving screen of variable size and brightness and an observer at an arbitrary viewing distance, the problem of locating the particular noise source that constitutes the bottle-neck for performance is not, in general, a simple one. For many problems of this type, however, a method of analysis which is at least conceptually simple, and usually operationally so, will be described. It is especially applicable to systems whose components show ideal performance.

The principle of the method may be outlined by considering a multistage electron multiplier. The signal-to-noise ratio coming out of such a multiplier is very closely the signal-to-noise ratio of the electron stream entering the multiplier. For any *normal* multiplier this is true. The noise contribution of the separate stages, especially for high gains per stage, is negligible. Suppose, however, that a stage somewhere in the middle of the multiplier turned out to have a gain less than unity. The signal-to-noise ratio out of the multiplier would still not be affected providing that the product of the gain of this stage and the total gain of all preceding stages was greater than unity. If this product were less than unity, the output current of the low gain stage would then determine the signal-to-noise ratio out of the multiplier. Or, generally, if one measured the electron currents coming into each stage of the multiplier the lowest of these currents would determine the final signal-to-noise ratio.

Similarly, the many transformations through which a picture element of the original scene goes before it emerges as a picture element in the brain of the final observer may be regarded as multiplierlike stages having various gains. The problem of finding the noise bottleneck in such a system becomes one of finding the smallest stream of particles or events representing the original picture element. An illustration of the method of analysis follows.

Fig. 13 shows schematically the elements of a system from scene through pickup tube and receiver to the final observer. Nine points of interest are marked out. At each point, the number of "events" representing an element of the original scene is indicated algebraically. At the bottom of the figure the numbers are arranged in a qualitative plot. Starting from an element of the original scene, N_o quanta are emitted per second in a lambert distribution. $N_o \sin^2 \phi_1$ of these quanta pass through the pickup tube lens whose half angle subtense at the scene is ϕ_1. At the photo surface of an ideal pickup tube, the $N_o \sin^2 \phi_1$ quanta are converted into a still fewer number ($\theta_1 N_o \sin^2 \phi_1$) of photoelectrons, θ_1 being the quantum yield of the surface. The pickup tube multiplies these signal electrons by the factor G_1, which is large enough that noise in the circuits connecting pickup tube and receiving tube may be neglected.

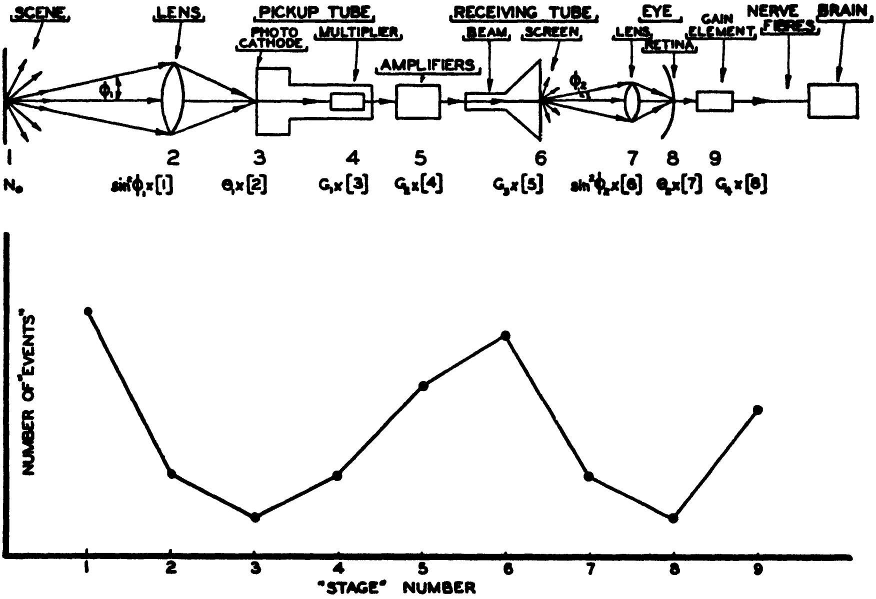


FIG. 13.—Chronicle of a picture element from the original scene to the brain of the final viewer.

These circuits further amplify the signal current by the factor G_2 so that the number of electrons in the receiving tube that represent the original element is about a thousand times greater than the number leaving the photocathode of the pickup tube. All of the gain factors are assumed to be constants characteristic of a *linear** system. The electrons in the receiving tube strike a luminescent screen and are converted into about 300 times as many light quanta. G_3 is this factor. If the quanta from the luminescent screen also follow a lambert distribution, the number passing through the eye lens is a small fraction, $\sin^2 \phi_2$, of the total number. Only a fraction θ_2 of those entering the eye are usefully absorbed at the retina. θ_2 is the quantum yield of the retina. Finally, some sort of gain element must amplify the effect of these absorbed quanta before they can generate the nerve pulses carrying information to the brain.

It is seen from the plot in Fig. 13 that there are two low points, one at the photosurface of the pickup tube and the other at the photosurface of the eye. If the pickup tube point is lower than that of the eye, the observer sees pickup tube noise on his receiving tube screen and the system performance is limited by the pickup tube. For the reverse order, the received picture is judged noise free and the system performance is limited by the eye. Technical economy would suggest a close match between limitation by eye and tube.

Fig. 13 shows also how the transition from a noisy to a noise-free picture can be made simply by decreasing the screen brightness at the receiving tube or by increasing the viewing distance. For the particular case in which the two brightnesses, angular subtenses and quantum yields are made equal, the pickup tube becomes in fact a transposed eye.

IX. INTELLIGENCE VS. BANDWIDTH AND SIGNAL-TO-NOISE RATIO

In the introduction to this paper, the phrase "bits of information" was used to characterize the intelligence transmitted by a pickup device. This usage was not by way of popularizing otherwise technical verbiage but rather to emphasize the finite character of the intelligence. That is, only a finite number of spatially separable elements and only a finite number of half tone steps per element are recognizable in a picture having a specified signal-to-noise ratio. In fact, one can easily count the total number of possible pictures that can be compounded out of n separate elements and m half tone steps per element. That number is m^n. On the other hand, the capacity of a system, having a bandwidth Δf and a signal-to-noise ratio R, for transmitting intelligence may also be measured by the total number of different pictures it can transmit. This number by the same reasoning is $\left(\frac{R}{k}\right)^{2T\Delta f}$ where k is the threshold signal-to-noise ratio and T is the time for one picture. Thus for each of the $2T\Delta f$ separable elements of time assigned to a picture there are R/k distinguishable values of signal amplitude that may be selected. Let the desired number of pictures be set equal to the number of different pictures that the system can transmit:

---

* For a nonlinear system, the *visibility* of noise on the receiving tube screen is proportional to the gamma of the system. The gamma of the system does *not*, however, vary the *signal-to-noise ratio* and, therefore, the intelligence transmitted. These statements are not to be confused with the arrangement of low gamma transmitter and high gamma receiver, which is designed to minimize the effects of noise picked up in transmission. The present discussion assumes that such noise may be made negligible.

$$m^n = \left(\frac{R}{k}\right)^{2T\Delta f} \tag{9}$$

The solution of this equation for Δf yields:

$$\Delta f = \frac{m}{2T}\frac{\log m}{\log\left(\frac{R}{k}\right)} \tag{10}$$

Eq. (10) says that when the signal-to-noise ratio is just large enough for the discrimination of the desired number of half tone steps, namely, if $\frac{R}{k} = m$, the bandwidth has its customary value of half the number of picture elements per second. If, on the other hand, the signal-to-noise ratio is larger than it need be for half tone discrimination, the bandwidth may be reduced by the factor $\frac{\log m}{\log R/k}$.* A specifically designed mechanism is needed, however, to effect this reduction.

A final comment should be made concerning bandwidths in excess of that needed to transmit the intelligence content of a picture. An oversize bandwidth does no particular harm to the picture transmitted by an *ideal* device. The noise content of such a picture is set by the picture itself and not by the bandwidth of the associated circuits. An exaggerated example of this is the series of photographs in Fig. 5. Here the bandwidth was about 5 "megacycles" while the picture content, for some of the shorter exposures, needed only a few kilocycles to convey substantially all of its information. Of course, wider bandwidths make it more difficult to transmit a picture without picking up noise in the transmission comparable with noise in the original picture.

---

* This reciprocity of bandwidth and signal-to-noise ratio (or its logarithm) was pointed out to the writer several years ago (1944) by Dr. G. A. Morton of RCA Laboratories Division.

Another example, effectively, of extremely wide band transmission is found in the photographic process. Using a microscope one can see the separate photographic grains whose size is of the order of a micron. Under normal conditions of viewing, however, the eye can see only elements larger than about 25 microns. That is, the eye sets the effective bandwidth at less than 1% of the capacity of the film. In fact, it is of no great consequence whether the eye can see the limiting resolution of film or not. The intelligence contained in elements near the limiting resolution of film is relatively small by virtue of the inherently low signal-to-noise ratio of these elements. For this reason it is not sensible to compare without qualification the number of lines of a television picture, for which number of lines the signal-to-noise ratio may be as high as needed, with the limiting resolution of film where, by definition, the signal-to-noise ratio approaches zero. A valid comparison must depend on relatively subjective tests which in turn are a critical function of picture content. Such tests and analyses have been carried out by Baldwin[21] and, more recently, in an extensive and thorough investigation by Schade.[4] Schade has found, for example, that equal picture quality can be transmitted by a television system having a number of scanning lines equal to half the limiting resolution of a grainy film.

Oversize bandwidths, while not penalizing the performance of an ideal device, do deteriorate the rendition of blacks and greys by a device, like the orthicon or iconoscope, for which amplifier noise is the limiting noise source.

X. CONCLUDING REMARKS

Lest the original purpose of this discussion not be recognizable in its elaboration, it is here restated. Because the final limitations of pickup devices are clearly and simply set by the quantum nature of light, because these limitations have not been widely discussed, and because the particular devices in the new and rapidly developing field of television are of less interest for the mechanics of their operations than they are as markers in the approach to ideal performance, the emphasis throughout has been on the setting up of an absolute scale of performance according to which the many and diverse pickup devices can be oriented. This approach is particularly needed because the eye, photographic film, and television pickup tubes are being called upon more and more frequently to critically pass upon each other's performance.

REFERENCES

1. Sziklai, G. C., Ballard, R. C., and Schroeder, A. C. *Proc. Inst. Radio Engrs.* **35,** 862–870 (1947).
2. Rose, A. *Proc. Inst. Radio Engrs.* **30,** 295–300 (1942).

3. DeVries, H. *Physica* **10,** 553–564 (1943).
4. Schade, O. H. *RCA Rev.* **9,** 5–37 (1948).
5. Cobb, P. W., and Moss, F. K. *J. Franklin Inst.* **205,** 831–847 (1928).
6. Connor, J. P., and Ganoung, R. E. *J. Opt. Soc. Amer.* **25,** 287–294 (1935).
7. Blackwell, H. R. *J. Opt. Soc. Amer.* **36,** 624–643 (1946).
8. Reeves, P. *Psychol. Rev.* **205,** 831 (1928).
9. Rose, A. *J. Opt. Sci. Amer.* **38,** 196–208 (1948).
10. Hecht, S. *J. Opt. Soc. Amer.* **32,** 42–49 (1942).
11. Jones, L. A., and Higgins, G. C. *J. Opt. Soc. Amer.* **36,** 203–227 (1946).
12. Zworykin, V. K., and Morton, G. A. Television. Wiley, New York, 1940.
13. Zworykin, V. K., and Ramberg, E. G. Section 15, part II, Electrical Engineer's Handbook. Edited by H. Pender and K. McIlwain, 4th ed. Wiley, New York, in press.
14. Farnsworth, P. T. *J. Franklin Inst.* **218,** 411–444 (1934).
15. Larson, C. C., and Gardner, B. C. *Electronics* **12,** 24–27 (Oct., 1939).
16. Zworykin, V. K., Morton, G. A., and Flory, L. E. *Proc. Inst. Radio Engrs.* **25,** 1071–1092 (1937).
17. Iams, H. A., Morton, G. A., and Zworykin, V. K. *Proc. Inst. Radio Engrs.* **27,** 541–547 (Sept., 1939).
18. Rose, A., and Iams, H. A. *RCA Rev.* **4,** 186–199 (1939).
19. Rose, A., Weimer, P. K., and Law, H. B. *Proc. Inst. Radio Engrs.* **34,** 424–432 (1946).
20. Rose, A. *J. Soc. Motion Picture Engrs.* **47,** 273–294 (1946).
21. Baldwin, M. W. *Proc. Inst. Radio Engrs.* **28,** 458–468 (1940).

The Deflection of Beams of Charged Particles

R. G. E. HUTTER

Sylvania Electric Products, Inc., Bayside, New York

CONTENTS

Page
I. Introduction 167
II. Small-Angle Deflection 168
  1. Field Distributions 170
  2. Equations of Motion 175
  3. Crossed Superimposed Two-Dimensional Electric Deflection Fields 176
  4. Single Three-Dimensional Electric Deflection Field 185
  5. Crossed Unbalanced Electric Deflection Fields 188
  6. Crossed Magnetic Deflection Fields 190
  7. Correction of Spot and Pattern Distortion 195
  8. Ion Traps and Linear Mass Spectrometers 199
III. Large-Angle Deflection 200
  1. Motion of Particles in Systems with Arbitrarily Curved Axes 201
  2. A Crossed Field Mass Spectrometer with a Radial Electric Field 205
  3. A Crossed Field Mass Spectrometer with a Constant Electric Field 208
  4. The Determination of Curved Optical Axes 210
  5. Two-Directional Focusing with Deflection Type Fields 212
  6. Purely Electric Deflection Fields 215
  7. Purely Magnetic Deflection Fields 216
  References 218

I. INTRODUCTION

The application of electron optical principles to problems of electron and ion beam formation has become a well-established procedure. A general theory of electron and ion lenses exists which may be applied to any kind of special field having rotational symmetry about an axis.

Problems of beam deflection have not, however, received the same systematic kind of treatment. Many different methods have been used in dealing with the motion of charged particles through such fields which perform the function of electron and ion prisms.

It is the purpose of this paper to give a unified presentation of the theory of prisms using electron optical methods throughout and to discuss a number of designs of prisms having improved properties.

Electric and/or magnetic deflection type fields are used in cathode ray and television tubes and ion traps for such tubes, beam-deflection amplifiers, scanning type electron microscopes, mass spectrometers, microanalyzers, beta-ray spectrometers, and cyclotrons, to name only the more important devices. In the first four instruments, the angle through which the beam is deflected is considerably smaller than that in the other four devices.

It will be shown that beam-focusing effects are invariably associated with the deflection which is produced by simple fields and that these focusing effects increase with the angle of deflection. Such effects are ordinarily undesirable in small-angle deflection devices while they are extremely useful in instruments employing large-angle deflection. It is not necessary to point out further differences between the various instruments since they do not result in different methods of theoretical procedures.

The most important electron-beam device employing small-angle deflection is the cathode ray tube. The terminology used in the first chapter of this paper is chosen to apply directly to the cathode ray tube. The application of the small angle deflection theory to devices like ion traps, linear mass spectrometers, and scanning-type electron microscopes will be obvious. The theory of large angle deflection, outlined in section III, will apply to mass spectrometers, beta-ray spectrometers, and microanalyzers.

II. Small-Angle Deflection

The two simplest types of deflection fields are the uniform electric field and the uniform magnetic field.* Fig. 1 shows a pair of parallel metal plates symmetrically displaced about the axis along which an electron beam travels with a speed corresponding to a voltage of V volts. The electrodes are connected to a potential source which brings the upper plate to the potential $(\phi + \frac{1}{2}V)$ volts and the lower plate to $(\phi - \frac{1}{2}V)$ volts. The electrostatic field between the plates will be nearly uniform over a large area if the linear dimensions of the plates are large compared with the distance between the plates.

Fig. 2 shows an arrangement of pole pieces which produces a nearly uniform magnetic field if the linear dimensions of the pole pieces are large compared with the distance between them. The pole pieces are energized by coils of an electromagnet through which a current of I amperes passes. The magnitude of the displacement of the beam at the target (e.g., a fluorescent screen) may readily be computed. If the results are applied to actual deflection fields, large discrepancies between calculated and measured values are observed. This is due to the fact that actual fields are nonuniform. If now the mathematical methods are improved so that fringe fields can be taken into account, it is found that a displacement is not strictly a linear function of the deflection voltage or current and that it depends on the initial position and direction of motion of the electron before entering the deflection fields. The explanation of the so-called pattern and spot distortions, observed in cathode ray and

---

* Deflection fields as used in cathode ray and television tubes are varying periodically in time. Ordinarily these fields may be considered as static since the fields do not change appreciably during a time interval equal to the transit time of the electrons passing through the field region.

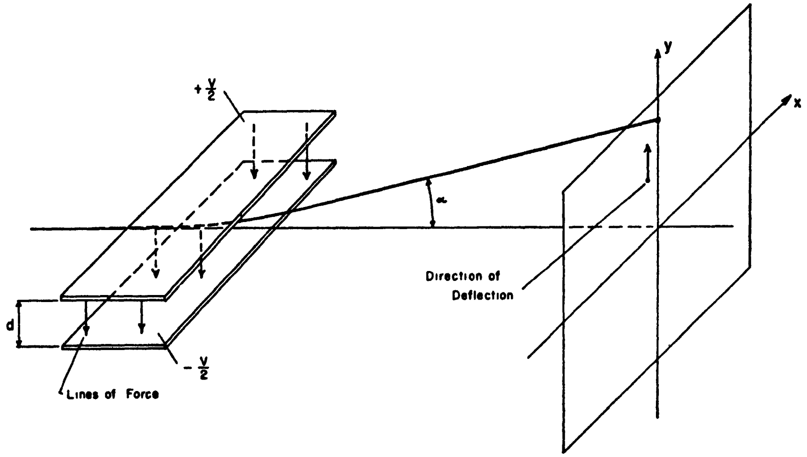


Fig. 1.—Electrostatic deflection system consisting of two parallel plates.

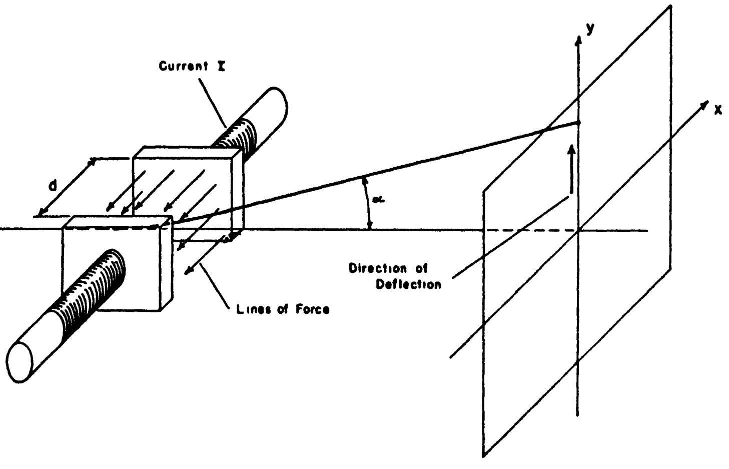


Fig. 2.—Magnetic deflection system consisting of two parallel pole pieces energized by an electromagnet.

television tubes, may be based on the two mentioned properties. A pattern is produced by the simultaneous action of two mutually perpendicular deflection fields varying independently in strength. If the deflection voltages are proportional to the parameters of the information which is to be made visual on the screen of the cathode ray tube, the image will not give a true graphical presentation if the deflection is a nonlinear function of the deflection voltage. Spot distortion is most objectionable in television tubes, since it results in a loss of resolution in the outer portions of the television image. Different parts of the electron beam have different initial conditions and hence are deflected by different amounts. Spot distortion would exist even if the target surface were spherical with a radius equal to the distance from the center of the deflection system to the screen.

In order to derive mathematical expressions for the defocusing and distortion effects, the path of an electron through the electromagnetic deflection field must be determined. The field distribution of deflection fields used in practice is, in general, too complicated to permit an integration of the equation of motion in terms of elementary functions. In general, therefore, methods of successive approximation must be applied. Such a procedure was followed by Picht and Himpan,[1] Wendt,[2] and Hutter[3,4] for various kinds of deflection fields. In order to solve the equations of motion of an electron through electromagnetic fields by such methods it is necessary to develop series expressions for the field-strength distribution of the fields.

1. *Field Distributions*

A vector field distribution $\vec{S}(x, y, z)$ which satisfies the conditions that

$$\operatorname{curl} \vec{S} = 0, \qquad \operatorname{div} \vec{S} = 0, \tag{1}$$

may be expressed as the negative gradient of a scalar potential function $\psi(x, y, z)$, i.e.,

$$\vec{S} = -\operatorname{grad} \psi. \tag{2}$$

This function satisfies LaPlace's equation,

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = 0. \tag{3}$$

Assuming that such a potential distribution is symmetrical about the yz plane, i.e.,

$$\psi(x, y, z) = \psi(-x, y, z), \tag{4}$$

an infinite series for ψ in powers of x and y would be of the form:

$$\psi(x, y, z) = \sum_{m,n}^{0\ldots\infty} S_{2n,m}(z) \cdot x^{2n} \cdot y^{m}. \tag{5}$$

Substituting ψ in eq. (3) yields a recurrence formula for the coefficients $S_{2n,m}$. If use is made of such a formula and the coefficients $S_{2n,0}$ and $S_{2n,1}$ are renamed

$$S_{2n,0} = \psi_{2n}, \qquad S_{2n,1} = P_{2n}, \tag{6}$$

the following series for ψ is obtained:

$$\begin{aligned}\psi(x, y, z) = {} & \psi_0(z) - P_0(z) \cdot y + \psi_2(z)x^2 - \tfrac{1}{2}(\psi_0''(z) + 2\psi_2(z))y^2 - \\ & - P_2(z)x^2y + \tfrac{1}{6}(P_0''(z) + 2P_2(z))y^3 + \psi_4(z)x^4 - \\ & - \tfrac{1}{2}(\psi_2''(z) + 12\psi_4(z))x^2y^2 + \tfrac{1}{24}(\psi_0^{(IV)}(z) + 2\psi_2''(z) - \\ & - 4\psi_4(z))y^4 - P_4(z) \cdot x^4y + \tfrac{1}{6}(P_2''(z) + 12P_4(z))x^2y^3 - \\ & - \tfrac{1}{120}(P_0^{(IV)}(z) + 4P_2''(z) + 24P_4(z))y^5 + \cdots .\end{aligned} \tag{7}$$

If x and y are interchanged in eq. (7), a field is obtained which is symmetrical about the xz-plane. The potential distribution of a field resulting from the superposition of two fields, one symmetrical about the yz-plane, the other symmetrical about the xz-plane, may then be written as

$$\begin{aligned}\psi(x, y, z) = {} & \psi_{01} + \psi_{02} - P_{01}y - P_{02}x - [\tfrac{1}{2}(\psi_{01}'' + 2\psi_{21}) - \psi_{22}]y^2 - \\ & - [\tfrac{1}{2}(\psi_{02}'' + 2\psi_{22}) - \psi_{21}]x^2 - P_{21}x^2y - P_{22}xy^2 + \\ & + \tfrac{1}{6}(P_{01}'' + 2P_{21})y^3 + \tfrac{1}{6}(P_{02}'' + 2P_{22})x^3 + [\psi_{42} + \\ & + \tfrac{1}{24}(\psi_{01}^{(IV)} + 2\psi_{21}'' - 4\psi_{41})]y^4 - [\tfrac{1}{2}(\psi_{21}'' + 12\psi_{41}) + \\ & + \tfrac{1}{2}(\psi_{22}'' + 12\psi_{42})]x^2y^2 + [\psi_{41} + \tfrac{1}{24}(\psi_{02}^{(IV)} + 2\psi_{22}'' - \\ & - 4\psi_{42})]x^4 - P_{41}x^4y - P_{42}xy^4 + \tfrac{1}{6}(P_{21}'' + 12P_{41})x^2y^3 + \\ & + \tfrac{1}{6}(P_{22}'' + 12P_{42})x^3y^2 - \tfrac{1}{120}(P_{01}^{(IV)} + 4P_{21}'' + \\ & + 24P_{41})y^5 - \tfrac{1}{120}(P_{02}^{(IV)} + 4P_{22}'' + 24P_{42})x^5 + \\ & + \cdots .\end{aligned} \tag{8}$$

where $\psi_{2n,1}$ and $P_{2n,1}$ are the functions determining one field while $\psi_{2n,2}$ and $P_{2n,2}$ are the functions determining the other field.

The relations derived so far apply to all vector fields and hence may also be used to describe electric and magnetic fields. Electric fields are ordinarily produced by electrodes connected to potential sources while magnetic fields are created by pole pieces of magnetic material or current carrying conductors. The coefficients of the infinite series are uniquely determined by the boundary conditions, i.e., by the shape, location of the electrodes, or current conductors and their potentials and currents. If the potentials for two such fields are computed separately as if existing alone in free space, it is found that, in general, the true potential distribution of the combined system is not the sum of the potentials of the two separate systems.

The electrodes and/or current conductors, producing one field, distort the field produced by the second set of electrodes and/or current conductors. Induced charges and/or currents cause the field of one system to vary if the potentials and/or currents of the other one are changed. This effect is termed "cross-talk." Since it cannot be tolerated in cathode ray and television tubes, one tries to achieve in practice the independence of two deflection systems by separating them as much as possible. The potential of the combined system may therefore be obtained by adding the potentials of the two deflection systems which are computed separately.

The field strength components of a vector field described by the potential distribution eq. (8) are given by:

$$\begin{aligned}
-\frac{\partial\psi}{\partial x} = {} & P_{02} + [(\psi_{02}'' + 2\psi_{22}) - 2\psi_{21}]x + 2P_{21}xy + P_{22}y^2 - \\
& - \tfrac{1}{2}(P_{02}'' + 2P_{22})x^2 + [\psi_{21}'' + 12\psi_{41} + \psi_{22}'' + 12\psi_{42}]xy^2 - \\
& - [4\psi_{41} + \tfrac{1}{6}(\psi_{02}^{(IV)} + 2\psi_{22}'' - 4\psi_{42})]x^3 + 4P_{41}x^3y + P_{42}y^4 - \\
& - \tfrac{1}{3}(P_{21}'' + 12P_{41})xy^3 - \tfrac{1}{2}(P_{22}'' + 12P_{42})x^2y^2 + \\
& + \tfrac{1}{24}(P_{02}^{(IV)} + 4P_{22}'' + 24P_{42})x^4 + \cdots,
\end{aligned}$$

$$\left.\begin{aligned}
-\frac{\partial\psi}{\partial y} = {} & P_{01} + [(\psi_{01}'' + 2\psi_{21}) - 2\psi_{22}]y + P_{21}x^2 + 2P_{22}xy - \\
& - \tfrac{1}{2}(P_{01}'' + 2P_{21})y^2 - [4\psi_{42} + \tfrac{1}{6}(\psi_{01}^{(IV)} + 2\psi_{21}'' - \\
& - 4\psi_{41})]y^3 - [(\psi_{21}'' + 12\psi_{41}) + (\psi_{22}'' + 12\psi_{42})]x^2y + \\
& + P_{41}x^4 + 4P_{42}xy^3 - \tfrac{1}{2}(P_{21}'' + 12P_{41})x^2y^2 - \\
& - \tfrac{1}{3}(P_{22}'' + 12P_{42})x^3y + \tfrac{1}{24}(P_{01}^{(IV)} + 4P_{21}'' + \\
& + 24P_{41})y^4 - + \cdots, \\
-\frac{\partial\psi}{\partial z} = {} & -\psi_{01}' - \psi_{02}' + P_{01}'y + P_{02}'x + \left[\frac{1}{2}(\psi_{01}''' + 2\psi_{21}') - \right. \\
& \left. - \psi_{22}'\right] y^2 + [\tfrac{1}{2}(\psi_{02}''' + 2\psi_{22}') - \psi_{21}']x^2 + P_{21}'x^2y + \\
& + P_{22}'xy^2 - \tfrac{1}{6}(P_{01}''' + 2P_{21}')y^3 - \tfrac{1}{6}(P_{02}''' + 2F_{22}')x^3 - \\
& - [\psi_{42}' + \tfrac{1}{24}(\psi_{01}^{(V)} + 2\psi_{21}''' - 4\psi_{41}')]y^4 + [\tfrac{1}{2}(\psi_{21}''' + \\
& + 12\psi_{41}') + \tfrac{1}{2}(\psi_{22}''' + 12\psi_{42}')]x^2y^2 - [\psi_{41}' + \tfrac{1}{24}(\psi_{02}^{(V)} + \\
& + 2\psi_{22}''' - 4\psi_{42}')]x^4 - + \cdots.
\end{aligned}\right\} \tag{9}$$

The deflection and deflection defects of a system described by eqs. (8) or (9) have never been studied and any attempt to do so would result in formulas difficult to interpret. However, the following simpler types of deflection systems have been investigated:

(1) A single, balanced electrostatic deflection field produced by electrodes which are infinitely long in the x direction and which are parallel to each other. Fields of this type are called two-dimensional since the field quantities are functions of only two variables. A particular electrode system producing such a field is sketched in Fig. 3. It consists of a pair of parallel plates bent on one end along lines in the x direction to make equal and opposite angles with the (x, z) plane. In this case, $\psi_{01} = \phi_0 = \text{const}$, all other $\psi_{2n,1}$ and $\psi_{2n,2}$ are equal to zero, $P_{01} = E(z)$ all other $P_{2n,1}$ and $P_{2n,2}$ are equal to zero. The electrostatic potential φ is given by:

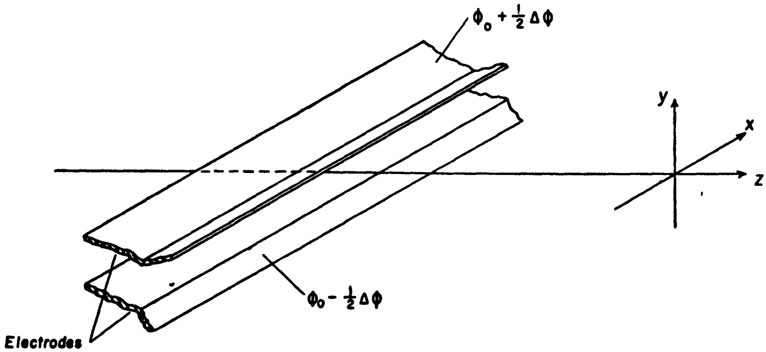

$\Phi_0 + \frac{1}{2}\Delta\Phi$
y
x
z
$\Phi_0 - \frac{1}{2}\Delta\Phi$
Electrodes

FIG. 3.—Single, balanced, two-dimensional electric deflection system. The electrode surfaces are generated by infinitely long, straight lines parallel to the x axis. The potential is a function of y and z only.

$$\varphi(x, y, z) = \phi_0 - E(z)y + \tfrac{1}{6}E''(z)y^3 - \tfrac{1}{120}E^{(\mathrm{IV})}(z)y^5 + \cdots . \quad (10)$$

It follows from eq. (10) that

$$\varphi(x, y, z) - \phi_0 = -\varphi(x, -y, z) - \phi_0. \quad (11)$$

Because of this property, the field is called a balanced deflection field.

(2) Two crossed, balanced electrostatic deflection fields of the type described under (1). It is assumed that the two functions $E_{01}(z)$ and $E_{02}(z)$ are essentially different from zero in two regions of z which do not overlap. The effects of the two fields may then be computed one after the other.

(3) Two crossed, balanced, overlapping elestrostatic deflection fields of the type described under (1). The two functions $E_{01}(z)$ and $E_{02}(z)$ may now be essentially different from zero for the same values of the axial coordinate z. In this case, the potential is given by:

$$\varphi(x, y, z) = \phi_0 - (E_{01}(z)y + E_{02}(z)x) + \tfrac{1}{6}(E_{01}''(z)y^3 + E_{02}''(z)x^3) - \\ - \tfrac{1}{120}(E_{01}^{(IV)}(z)y^5 + E_{02}^{(IV)}(z)x^5) - + \cdots . \quad (12)$$

(4) Two crossed, unbalanced electrostatic deflection fields, each being produced by a pair of electrodes infinitely long in the x or y direction. The electrodes of each pair are parallel to each other. The potential distribution may be written

$$\varphi(x, y, z) = \phi_{01} + \phi_{02} - E_{01}y - E_{02}x - \tfrac{1}{2}\phi_{01}''y^2 - \tfrac{1}{2}\phi_{02}''x^2 + \\ + \tfrac{1}{6}E_{01}''y^3 + \tfrac{1}{6}E_{02}''x^3 + \tfrac{1}{24}\phi_{01}^{(IV)}y^4 + \tfrac{1}{24}\phi_{02}^{(IV)}x^4 - \\ - \tfrac{1}{120}E_{01}^{(IV)}y^5 - \tfrac{1}{120}E_{02}^{(IV)}x^5 + \cdots . \quad (13)$$

For later purposes we write

$$\phi_{01}(z) + \phi_{02}(z) = \phi_m + \phi_{m1}\phi_1(z) + \phi_{m2}\phi_2(z). \quad (14)$$

(5) Single, balanced electrostatic deflection fields deflecting in the y direction. The field strength is assumed to vary along the x-axis.

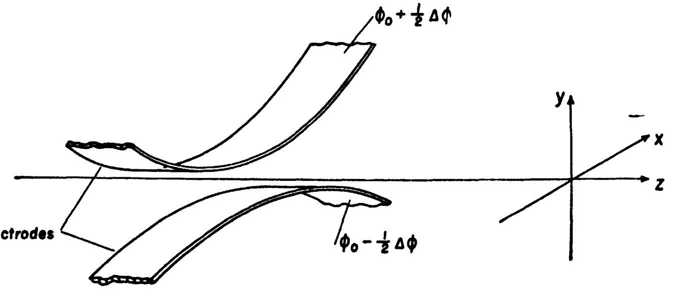

$\phi_0 + \frac{1}{2}\Delta\phi$
y
x
z
Electrodes
$\phi_0 - \frac{1}{2}\Delta\phi$

FIG. 4.—Single, balanced, three-dimensional electric deflection system. The potential is a function of all coordinates x, y, and z.

The only requirement for electrodes producing such a field is that the two surfaces are symmetrical about the (y, z)-and (x, z)-planes. Fig. 4 shows a particular form of such a system. It can be seen that the field strength will vary in the x direction since the distance between the two electrodes is a function of x and the potential difference between the electrodes is constant. The potential distribution becomes:

$$\varphi(x, y, z) = \phi_0 - E_0y + \tfrac{1}{6}(E_0'' + 2E_2)y^3 - E_2x^2y + \\ + \tfrac{1}{6}(E_2'' + 12E_4)x^2y^3 - \tfrac{1}{120}(E_0^{(IV)} + 4E_2'' + 24E_4)y^5 - \\ - E_4x^4y + - \cdots . \quad (15)$$

(6) Two crossed magnetic deflection fields. Let $P_{2n,1} = H_{2n,1}$ and $P_{2n,2} = -H_{2n,2}$ and $\psi_{2n,1} = \psi_{2n,2} = 0$. The field strength components become:

$$\left.\begin{aligned} H_x &= -H_{02} + (H_{22} + \tfrac{1}{2}H_{02}'')x^2 + 2H_{21}xy - H_{22}y^2 - \\ &\quad - (H_{42} + \tfrac{1}{6}H_{22}'' + \tfrac{1}{24}H_{02}^{(\mathrm{IV})})x^4 + 4H_{41}x^3y + \\ &\quad + (6H_{42} + \tfrac{1}{2}H_{22}'')x^2y^2 - (4H_{41} + \tfrac{1}{3}H_{21}'')xy^3 - \\ &\qquad\qquad - H_{42}y^4 + \cdots, \\ H_y &= H_{01} + H_{21}x^2 - 2H_{22}xy - (H_{21} + \tfrac{1}{2}H_{01}'')y^2 + H_{41}x^4 + \\ &\quad + (4H_{42} + \tfrac{1}{3}H_{22}'')x^3y - (6H_{41} + \tfrac{1}{2}H_{21}'')x^2y^2 - 4H_{42}xy^3 + \\ &\qquad + (H_{41} + \tfrac{1}{6}H_{21}'' + \tfrac{1}{24}H_{01}^{(\mathrm{IV})})y^4 + \cdots, \\ H_z &= -H_{02}'x + H_{01}'y - H_{22}'xy^2 + H_{21}'x^2y + (\tfrac{1}{3}H_{22}' + \\ &\qquad + \tfrac{1}{6}H_{02}''')x^3 - (\tfrac{1}{3}H_{21}' + \tfrac{1}{6}H_{01}''')y^3 + \cdots. \end{aligned}\right\} \quad (16)$$

(7) A single magnetic deflection field. The pole pieces shown in Fig. 2 would actually produce such a field. If $H_{2n,1}$ is equal to zero, eq. (16) goes over into

$$\left.\begin{aligned} H_x &= -H_{02} + (H_{22} + \tfrac{1}{2}H_{02}'')x^2 - H_{22}y^2 - (H_{42} + \tfrac{1}{6}H_{22}'' + \\ &\quad + \tfrac{1}{24}H_{02}^{(\mathrm{IV})})x^4 + (6H_{42} + \tfrac{1}{2}H_{22}'')x^2y^2 - H_{42}y^2 + \cdots, \\ H_y &= 2H_{22}xy - 4H_{42}xy^3 + (4H_{42} + \tfrac{1}{3}H_{22}'')x^3y + \cdots, \\ H_z &= -H_{02}'x - H_{22}'xy^2 + (\tfrac{1}{3}H_{22}' + \tfrac{1}{6}H_{02}''')x^3 + \cdots. \end{aligned}\right\} \quad (17)$$

2. Equations of Motion

The equations of motion of an electron in an electromagnetic field are written advantageously in the following form:

$$\left.\begin{aligned} \frac{d}{dz}\left(\frac{\partial F}{\partial x'}\right) - \frac{\partial F}{\partial x} = 0, \\ \frac{d}{dz}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y} = 0, \end{aligned}\right\} \quad (18)$$

where F is given by

$$F(x, y, z, x', y') = \sqrt{\frac{2e}{m}\,\varphi(x, y, z)}\,\sqrt{1 + x'^2 + y'^2} - \\ - \frac{e}{m}(A_x x' + A_y y' + A_z); \quad (19)$$

A_x, A_y, A_z are the components of the magnetic vector potential which is related to the magnetic field strength by

$$\vec{H} = \operatorname{curl} \vec{A},$$

e and m are the charge and mass of the electron respectively. The primes indicates derivatives with respect to z. This expression may be expanded in a power series of powers of x, y, x' and y'. The integration of eq. (18) is then performed using well-known methods of successive approximation.

In addition to the cases of the crossed electrostatic and crossed magnetic deflection fields, which have been described in the literature, the case of a single electric deflection field for which the potential distribution is a function of all three coordinates (case 5) and the case of crossed unbalanced electrostatic deflection fields will be discussed. The latter type deflection field was treated theoretically by Cazalas[5] using, however, different methods. Solutions for the other three types of deflection fields (cases 1, 2, and 7) are then special cases of the expressions derived here.

3. Crossed Superimposed Two-Dimensional Electric Deflection Fields

The potential function is given by eq. (12). The Euler-Lagrange differential equations are given by:

$$\left.\begin{aligned}
&\frac{d}{dz}\left[x' - \frac{1}{2\phi_0}(E_{01}y + E_{02}x)x' - \frac{1}{8\phi_0{}^2}x'(E_{01}y + E_{02}x)^2 - \right.\\
&\quad\left. - \frac{1}{2}(x'^2 + y'^2)x'\right] = \frac{1}{2\phi_0}E_{02} + \frac{1}{4\phi_0}E_{02}''x^2 - \frac{1}{4\phi_0{}^2}(E_{01}y + \\
&\quad + E_{02}x)E_{02} + \frac{1}{24\phi_0{}^2}E_{02}(E_{01}''y^3 + E_{02}''x^3) + \\
&\quad + \frac{1}{8\phi_0{}^2}E_{02}''x^2(E_{01}y + E_{02}x) - \frac{3}{16\phi_0{}^3}(E_{01}y + E_{02}x)^2E_{02} - \\
&\quad - \frac{5}{32\phi_0{}^4}(E_{01}y + E_{02}x)^3 - \frac{1}{4\phi_0}(x'^2 + y'^2)E_{02} - \\
&\qquad\qquad - \frac{1}{8\phi_0{}^2}(x'^2 + y'^2)(E_{01}y + E_{02}x)E_{02},\\
&\frac{d}{dz}\left[y' - \frac{1}{2\phi_0}y'(E_{01}y + E_{02}x) - \frac{1}{8\phi_0{}^2}(E_{01}y + E_{02}x)^2y' - \right.\\
&\quad\left. - \frac{1}{2}(x'^2 + y'^2)y'\right] = - \frac{1}{2\phi_0}E_{01} + \frac{1}{4\phi_0}E_{01}''y^2 - \frac{1}{4\phi_0{}^2}(E_{01}y + \\
&\quad + E_{02}x)E_{02} + \frac{1}{24\phi_0{}^2}(E_{01}''y^3 + E_{02}''x^3)E_{01} + \\
&\quad + \frac{1}{8\phi_0{}^2}E_{01}''y^2(E_{01}y + E_{02}x) - \frac{3}{16\phi_0{}^3}(E_{01}y + E_{02}x)^2 - \\
&\quad - \frac{5}{32\phi_0{}^4}(E_{01}y + E_{02}x)^3E_{01} - \frac{1}{4\phi_0}(x'^2 + y'^2)E_{01} - \\
&\qquad\qquad - \frac{1}{8\phi_0{}^2}(x'^2 + y'^2)E_{01}(E_{01}y + E_{02}x).
\end{aligned}\right\} \qquad (20)$$

Here all terms up to the third order in x, y, x' and y' have been included.

The first step consists of the integration of the equations

$$x'' = - \frac{1}{2\phi_0}E_{02}, \qquad y'' = - \frac{1}{2\phi_0}E_{01}. \qquad (21)$$

The solutions of these equations will represent a good approximation to the actual electron path if

$$\left|\frac{1}{2}\frac{E_{01}}{\phi_0}y\right| \ll 1, \qquad \left|\frac{1}{2}\frac{E_{02}}{\phi_0}x\right| \ll 1. \tag{22}$$

This assumption means that the deflection is considered to be small in regions where the axial field-strength distribution is essentially different from zero. The solutions of eq. (21) are

$$x_g(z) = x_{iu} + x_{iu}'(z - z_i) + X(z), \qquad y_g(z) = y_{iu} + y_{iu}'(z - z_i) + Y(z), \tag{23}$$

where

$$\left.\begin{aligned} X(z) &= -\frac{1}{2\phi_0}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} E_{02}(u)du, \\ Y(z) &= -\frac{1}{2\phi_0}\int_{0}^{z} d\xi \int_{z_0}^{\xi} E_{01}(u)du. \end{aligned}\right\} \tag{24}$$

The quantities x_{iu}, y_{iu} and x_{iu}', y_{iu}' are the coordinates and slopes of an undeflected electron at the screen position $z = z_i$ respectively. In a plane $z = z_0$ the coordinates and slopes of such an electron may be determined by solving the equations:

$$\left.\begin{aligned} x_0 &= x_{iu} + x_{iu}'(z_0 - z_i), & y_0 &= y_{iu} + y_{iu}'(z_0 - z_i), \\ x_0' &= x_{iu}', & y_0' &= y_{iu}'. \end{aligned}\right\} \tag{25}$$

In order to obtain a better approximation to the actual electron path, the functions $x_g(z)$, $y_g(z)$ of eq. (19) are substituted in all terms of eq. (16) which are of higher order than the first. New solutions $x(z)$, $y(z)$ may then be determined by simple integrations. The differences $x - x_g$ and $y - y_g$ between the first order electron path and the next higher approximation are called Δx and Δy. The values of these quantities at $z = z_i$ are

$$\left.\begin{aligned} \Delta x_i &= \sum_{abcd}^{0\ldots\infty} A_{abcd}x_{iu}^a y_{iu}^b x_{iu}'^c y_{iu}'^d, \\ \Delta y_i &= \sum_{abcd}^{0\ldots\infty} B_{abcd}x_{iu}^a y_{iu}^b x_{iu}'^c y_{iu}'^d, \end{aligned}\right\} \tag{26}$$

where the coefficients B_{abcd} are given by

$$B_{0000} = \left\{\frac{1}{2}\int Y'^3dz + \frac{1}{4\phi_0}\int [(z - z_i)(E_{01}Y'^2 - E_{01}''Y^2) + \right. \tag{27}$$

$$+ 2E_{01}YY']dz + \frac{1}{24\phi_0^2}\int [(z - z_i)(6E_{01}^2Y - 4E_{01}E_{01}''Y^3 +$$

$$
\begin{aligned}
&+ 3E_{01}{}^2YY'^2) + 3E_{01}{}^2Y^2Y']dz + \frac{3}{16\phi_0{}^3}\int E_{01}{}^3Y^2(z - z_i)dz + \\
&+ \frac{5}{32\phi_0{}^4}\int E_{01}{}^4Y^3(z - z_i)dz\Big\} + \frac{1}{2}\int X'^2Ydz + \\
&+ \frac{1}{4\phi_0}\int [E_{01}X'^2(z - z_i) + 2E_{02}XY']dz + \frac{1}{24\phi_0{}^2}\int [(z - \\
&- z_i)(6E_{01}E_{02}X - E_{01}E_{02}''X^3 - 3E_{02}E_{01}''XY^2 + 3E_{01}{}^2YX'^2 + \\
&+ 3E_{01}E_{02}XX'^2 + 3E_{02}E_{01}XY'^2) + 6E_{02}E_{01}XYY' + \\
&+ 3E_{02}{}^2X^2Y']dz + \frac{3}{16\phi_0{}^3}\int (z - z_i)(2E_{02}E_{01}{}^2XY + \\
&+ E_{01}{}^2E_{02}X^2)dz + \frac{5}{32\phi_0{}^4}\int (z - z_i)(E_{02}{}^3E_{01}X^3 + \\
&\qquad\qquad + 3E_{01}{}^3E_{02}XY^2 + 3E_{02}{}^2E_{01}{}^2YX^2)dz,
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
B_{1000} = {}& \frac{1}{2\phi_0}\int E_{02}Y'dz + \frac{1}{8\phi_0{}^2}\int [(z - z_i)(E_{02}E_{01}X'^2 + E_{02}E_{01}Y'^2 + \\
&+ 2E_{02}E_{01} - E_{01}E_{02}''X^2 - E_{02}E_{01}''Y^2) + 2(E_{01}E_{02}YY' + \\
&+ E_{02}{}^2XY')]dz + \frac{3}{8\phi_0{}^3}\int (z - z_i)(E_{02}{}^2E_{01}X + E_{02}E_{01}{}^2Y)dz + \\
&+ \frac{15}{32\phi_0{}^4}\int (z - z_i)(E_{01}{}^3E_{02}Y^2 + 2E_{02}{}^2E_{01}{}^2XY + E_{01}E_{02}{}^2X^2)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0100} = {}& \Big\{\frac{1}{2\phi_0}\int [E_{01}Y' - E_{01}''Y(z - z_i)]dz + \frac{1}{8\phi_0{}^2}\int [(z - z_i)(2E_{01}{}^2 - \\
&- 4E_{01}E_{01}''Y^2 + E_{01}{}^2Y'^2) + 2E_{01}{}^2YY']dz + \frac{3}{8\phi_0{}^3}\int E_{01}{}^3Y(z - \\
&- z_i)dz + \frac{15}{32\phi_0{}^4}\int E_{01}{}^4Y^2(z - z_i)dz\Big\} + \frac{1}{8\phi_0{}^2}\int [(z - \\
&- z_i)(E_{01}{}^2X'^2 - 2E_{02}E_{01}''XY) + 2E_{01}E_{02}XY']dz + \\
&+ \frac{3}{8\phi_0{}^3}\int E_{02}E_{01}{}^2X(z - z_i)dz + \frac{15}{32\phi_0{}^4}\int (z - z_i)(E_{02}{}^2E_{01}{}^2X^2 + \\
&\qquad\qquad + 2E_{01}{}^3E_{02}XY)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0010} = {}& \int X'Y'dz + \frac{1}{2\phi_0}\int (z - z_i)(E_{01}X' + E_{02}Y')dz + \\
&+ \frac{1}{8\phi_0{}^2}\int [(z - z_i)^2(E_{02}E_{01}Y'^2 + E_{02}E_{01}X'^2 - E_{02}E_{01}''Y^2 - \\
&- E_{01}E_{02}''X^2 + 2E_{01}E_{02}) + 2(z - z_i)(E_{01}{}^2X' + E_{02}E_{01}XX' + \\
&+ E_{01}E_{02}YY' + E_{02}XY')]dz + \frac{3}{8\phi_0{}^3}\int (z - z_i)^2(E_{02}E_{01}{}^2Y + \\
&+ E_{01}E_{02}{}^2X)dz + \frac{15}{32\phi_0{}^4}\int (z - z_i)^2(E_{02}{}^3E_{01}X^2 + \\
&\qquad\qquad + 2E_{01}{}^2E_{02}{}^2XY + E_{01}{}^3E_{02}Y^2)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0001} = {} & \left\{ \frac{3}{2} \int Y'^2 dz + \frac{1}{2\phi_0} \int [E_{01}Y + 2(z - z_i)E_{01}Y' - \right. \qquad (27) \\
& - (z - z_i)^2 E_{01}''Y]dz + \frac{1}{8\phi_0^2} \int [E_{01}^2 Y^2 + 4(z - z_i)E_{01}^2 YY' + \\
& + (z - z_i)^2 (2E_{01}^2 - 4E_{01}E_{01}''Y^2 + E_{01}^2 Y'^2)]dz + \\
& \left. + \frac{3}{8\phi_0^3} \int E_{01}^3 Y(z - z_i)dz + \frac{15}{32\phi_0^4} \int E_{01}^4 Y^2 (z - z_i)^2 dz \right\} + \\
& + \frac{1}{2} \int X'^2 dz - \frac{1}{2\phi_0} \int E_{02}Xdz + \frac{1}{8\phi_0^2} \int [E_{02}^2 X^2 + 2E_{01}E_{02}XY + \\
& + 4(z - z_i)E_{02}E_{01}XY' + (z - z_i)^2 (E_{01}^2 X'^2 - E_{02}E_{01}''XY)]dz + \\
& + \frac{3}{8\phi_0^3} \int E_{02}E_{01}^2 X(z - z_i)^2 dz + \frac{15}{32\phi_0^4} \int (z - z_i)^2 (E_{02}^2 E_{01}^2 X^2 + \\
& \qquad + 2E_{01}^3 E_{02} XY)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{2000} = {} & \frac{1}{8\phi_0^2} \int (E_{02}Y' - E_{01}E_{02}''X(z - z_i))dz + \frac{3}{16\phi_0^3} \int E_{02}^2 E_{01}(z - \\
& - z_i)dz + \frac{15}{32\phi_0^4} \int (z - z_i)(E_{02}^2 E_{01}^2 Y + E_{02}^3 E_{01} X)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0200} = {} & \left\{ - \frac{1}{4\phi_0} \int E_{01}''(z - z_i)dz + \frac{1}{8\phi_0^2} \int [E_{01}^2 Y' - 4(z - \right. \\
& - z_i)E_{01}E_{01}''Y]dz + \frac{3}{16\phi_0^3} \int E_{01}^3 (z - z_i)dz + \\
& \left. + \frac{15}{32\phi_0^4} \int E_{01}^4 Y(z - z_i)dz \right\} - \frac{1}{8\phi_0^2} \int E_{02}E_{01}''X(z - z_i)dz + \\
& + \frac{15}{32\phi_0^4} \int E_{01}^3 E_{02} X(z - z_i)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0020} = {} & \left\{ \frac{1}{2} Y_i + \frac{1}{4\phi_0} \int E_{01}(z - z_i)dz + \frac{1}{8\phi_0^2} \int E_{01}^2 Y(z - z_i)dz \right\} + \\
& + \frac{1}{8\phi_0^2} \int [(z - z_i) \cdot E_{02}E_{01}X + (z - z_i)^2 (2E_{02}E_{01}X' + \\
& + E_{02}^2 Y') - (z - z_i)^3 E_{01}E_{02}''X]dz + \frac{3}{16\phi_0^3} \int E_{02}^2 E_{01}(z - \\
& z_i)^3 dz + \frac{15}{32\phi_0^4} \int (z - z_i)^3 (E_{02}^2 E_{01}^2 Y + E_{02}^3 E_{01} X)dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0002} = {} & \left\{ \frac{1}{2\phi_0} E_{01}(z_0)(z_i - z_0)^2 + \frac{3}{2} Y_i + \frac{1}{4\phi_0} \int [3(z - z_i)E_{01} - \right. \\
& - (z - z_i)^3 E_{01}'']dz + \frac{1}{8\phi_0^2} \int [3(z - z_i)E_{01}^2 Y + 3(z - z_1)E_{01}^2 Y' - \\
& - 4(z - z_i)^3 E_{01}E_{01}''Y]dz + \frac{3}{16\phi_0^3} \int E_{01}^3 (z - z_i)^3 dz +
\end{aligned}
$$

$$
\begin{aligned}
&\left. + \frac{15}{32\phi_0^4} \int E_{01}^4 Y(z - z_i)^3 dz \right\} + \frac{1}{8\phi_0^2} \int [(z - z_i) 3 E_{01} E_{02} X - \\
&- E_{02} E_{01}'' X (z - z_i)^3] dz + \frac{15}{32\phi_0^4} \int E_{01}^3 E_{02} X (z - z_i)^3 dz, \qquad (27)
\end{aligned}
$$

$$
\begin{aligned}
B_{1100} = {} & \frac{1}{4\phi_0^2} \int [E_{01} E_{02} Y' - E_{02} E_{01}'' Y (z - z_i)] dz + \frac{3}{8\phi_0^3} \int E_{02} E_{01}^2 (z - \\
& \qquad - z_i) dz + \frac{15}{16\phi_0^4} \int [(z - z_i)(E_{01}^3 E_{02} X + E_{02}^2 E_{01}^2)] dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0011} = {} & \frac{1}{2\phi_0} E_{02}(z_0)(z_i - z_0)^2 + X_i + \frac{1}{2\phi_0} \int E_{02}(z - z_i) dz + \\
& + \frac{1}{4\phi_0^2} \int [(z - z_i)(E_{02}^2 X + E_{01} E_{02} Y) + (z - z_i)^2 (E_{01}^2 X' + \\
& + 2 E_{02} E_{01} Y') - E_{02} E_{01}'' (z - z_i)^3 Y] dz + \frac{3}{8\phi_0^3} \int E_{02} E_{01}^2 (z - \\
& \qquad - z_i)^3 dz + \frac{15}{16\phi_0^4} \int (z - z_i)^3 (E_{02}^2 E_{01}^2 X + E_{01}^3 E_{02} Y) dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{1010} = {} & \frac{1}{4\phi_0^2} \int [(z - z_i)(E_{02} E_{01} X' + E_{02}^2 Y') - (z - z_i)^2 E_{01} E_{02}'' X] dz + \\
& + \frac{3}{8\phi_0^3} \int E_{02}^2 E_{01} (z - z_i)^2 dz + \frac{15}{16\phi_0^4} \int (z - z_i)^2 (E_{02}^2 E_{01}^2 Y + \\
& \qquad + E_{02}^3 E_{01} X) dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{0101} = {} & \left\{ - \frac{1}{2\phi_0} E_{01}(z_0)(z_i - z_0) + \frac{1}{2\phi_0} \int [E_{01} - E_{01}'' (z - z_i)^2] dz + \right. \\
& + \frac{1}{4\phi_0^2} \int [E_{01}^2 Y + 2(z - z_i) E_{01}^2 Y' - 4(z - z_i)^2 E_{01} E_{01}'' Y] dz + \\
& \left. + \frac{3}{8\phi_0^3} \int E_{01}^3 (z - z_i)^2 dz + \frac{15}{16\phi_0^4} \int E_{01}^4 Y (z - z_i)^2 dz \right\} + \\
& + \frac{1}{4\phi_0^2} \int [E_{01} E_{02} X - (z - z_i)^2 E_{01}'' E_{02} X] dz - \\
& \qquad - \frac{15}{16\phi_0^4} \int E_{01}^3 E_{02} X (z - z_i)^2 dz,
\end{aligned}
$$

$$
\begin{aligned}
B_{1001} = {} & - \frac{1}{2\phi_0} E_{02}(z_0)(z_i - z_0) + \frac{1}{2\phi_0} \int E_{02} dz + \frac{1}{4\phi_0^2} \int [E_{02}^2 X + \\
& + E_{01} E_{02} Y + (z - z_i) E_{01} E_{02} Y' - (z - z_i)^2 E_{02} E_{01}'' Y] dz + \\
& + \frac{3}{8\phi_0^3} \int E_{02} E_{01}^2 (z - z_i)^2 dz + \frac{15}{32\phi_0^4} \int (z - z_i)^2 (E_{02}^2 E_{01}^2 X + \\
& \qquad + E_{01}^3 E_{02} Y) dz,
\end{aligned}
$$

$$B_{0110} = \frac{1}{4\phi_0^2} \int [(z - z_i)(E_{01}^2X' + E_{01}E_{02}Y') - (z - z_i)^2E_{02}E_{01}''Y]dz +$$

$$+ \frac{3}{8\phi_0^3} \int E_{02}E_{01}^2(z - z_i)^2dz + \frac{15}{16\phi_0^4} \int (z - z_i)^2(E_{02}^2E_{01}^2X +$$

$$+ E_{01}^3E_{02}Y)dz, \qquad (27)$$

$$B_{3000} = -\frac{1}{24\phi_0^2} \int E_{01}E_{02}''(z - z_i)dz + \frac{5}{32\phi_0^4} \int E_{02}^3E_{01}(z - z_i)dz,$$

$$B_{0300} = \left\{-\frac{1}{6\phi_0^2} \int E_{01}E_{01}''(z - z_i)dz + \frac{5}{32\phi_0^4} \int E_{01}^4(z - z_i)dz\right\},$$

$$B_{0030} = \frac{1}{24\phi_0^2} \int [3(z - z_i)^2E_{02}E_{01}'' - (z - z_i)^4E_{01}E_{02}'']dz +$$

$$+ \frac{5}{32\phi_0^4} \int E_{02}^3E_{01}(z - z_i)^4dz,$$

$$B_{0003} = \left\{-\frac{1}{8\phi_0^2} E_{01}^2(z_0)(z_i - z_0)^3 + \frac{1}{12\phi_0^2} \int [3(z - z_i)^2E_{01}^2 -\right.$$

$$\left. - 2(z - z_i)^4E_{01}E_{01}'']dz + \frac{5}{32\phi_0^4} \int E_{01}^4(z - z_i)^4dz\right\},$$

$$B_{2100} = \frac{15}{32\phi_0^4} \int E_{02}^2E_{01}^2(z - z_i)dz,$$

$$B_{1200} = -\frac{1}{8\phi_0^2} \int E_{02}E_{01}''(z - z_i)dz + \frac{15}{32\phi_0^4} \int E_{01}^3E_{02}(z - z_i)dz,$$

$$B_{2010} = -\frac{1}{8\phi_0^2} \int E_{01}E_{02}''(z - z\ )^2dz + \frac{15}{32\phi_0^4} \int E_{02}^3E_{01}(z - z_i)^2dz,$$

$$B_{2001} = -\frac{1}{8\phi_0^2} E_{02}^2\ (z_0)(z_i - z_0) + \frac{1}{8\phi_0^2} \int E_{02}^2dz +$$

$$+ \frac{15}{32\phi_0^4} \int E_{02}^2E_{01}^2(z - z_i)^2dz,$$

$$B_{0210} = -\frac{1}{8\phi_0^2} \int E_{02}E_{01}''(z - z_i)^2dz + \frac{15}{32\phi_0^4} \int E_{01}^3E_{02}(z - z_i)^2dz,$$

$$B_{0201} = \left\{-\frac{1}{8\phi_0^2} E_{01}^2(z_0)(z_i - z_0) + \frac{1}{8\phi_0^2} \int [E_{01}^2 - 4(z -\right.$$

$$\left. - z_i)^2E_{01}E_{01}'']dz + \frac{15}{32\phi_0^4} \int E_{01}^4(z - z_i)^2dz\right\},$$

$$B_{1101} = \frac{1}{4\phi_0^2} E_{01}(z_0)E_{02}(z_0)(z_i - z_0) + \frac{1}{4\phi_0^2} \int [E_{01}E_{02} -$$

$$- E_{02}E_{01}''(z - z_i)^2]dz + \frac{15}{16\phi_0^4} \int E_{01}^3E_{02}(z - z_i)^2dz,$$

$$
\begin{aligned}
B_{1110} &= \frac{15}{16\phi_0^4}\int E_{02}^2E_{01}^2(z-z_\iota)^2dz, \qquad (27)\\
B_{1020} &= \frac{1}{8\phi_0^2}\int [E_{02}E_{01}(z-z_\iota) - E_{01}E_{02}''(z-z_\iota)^3]dz + \\
&\qquad + \frac{15}{32\phi_0^4}\int E_{02}^3E_{01}(z-z_\iota)^3dz,\\
B_{0120} &= \left\{\frac{1}{8\phi_0^2}\int E_{01}^2(z-z_\iota)dz\right\} + \frac{15}{32\phi_0^4}\int E_{02}^2E_{01}^2(z-z_\iota)^3dz,\\
B_{1002} &= -\frac{1}{4\phi_0^2}E_{01}(z_0)E_{02}(z_0)(z_\iota-z_0)^2 + \frac{1}{8\phi_0^2}\int [3E_{02}E_{01}(z-z_\iota) - \\
&\qquad - E_{02}E_{01}''(z-z_\iota)^3]dz + \frac{15}{32\phi_0^4}\int E_{01}^3E_{02}(z-z_\iota)^3dz,\\
B_{0102} &= \left\{\frac{1}{4\phi_0^2}E_{01}^2(z_0)(z_\iota-z_0)^2 + \frac{1}{8\phi_0^2}\int [3E_{01}^2(z-z_\iota) - \right.\\
&\qquad \left. - 4(z-z_\iota)^3E_{01}E_{01}'']dz + \frac{15}{32\phi_0^4}\int E_{01}^4(z-z_\iota)^3dz\right\},\\
B_{1011} &= \frac{1}{4\phi_0^2}E_{02}^2(z_0)(z_\iota-z_0)^2 + \frac{1}{4\phi_0^2}\int E_{02}^2(z-z_\iota)dz + \\
&\qquad + \frac{15}{16\phi_0^4}\int E_{02}^2E_{01}^2(z-z_\iota)^3dz,\\
B_{0111} &= -\frac{1}{4\phi_0^2}E_{01}(z_0)E_{02}(z_0)(z_\iota-z_0)^2 + \frac{1}{4\phi_0^2}\int [E_{01}E_{02}(z-z_\iota) - \\
&\qquad - E_{02}E_{01}''(z-z_\iota)^3]dz + \frac{15}{16\phi_0^4}\int E_{01}^3E_{02}(z-z_\iota)^3dz,\\
B_{0021} &= \left\{+\frac{1}{8\phi_0^2}\int E_{01}^2(z-z_\iota)^2dz\right\} - \frac{1}{8\phi_0^2}E_{02}^2(z_0)(z_\iota-z_0)^3 + \\
&\qquad + \frac{1}{8\phi_0^2}\int E_{02}^2(z-z_\iota)^2dz + \frac{15}{32\phi_0^4}\int E_{02}^2E_{01}^2(z-z_\iota)^4dz,\\
B_{0012} &= \frac{1}{4\phi_0^2}E_{02}(z_0)E_{01}(z_0)(z_\iota-z_0)^3 + \frac{1}{8\phi_0^2}\int [3(z-z_\iota)^2E_{02}E_{01} - \\
&\qquad - E_{02}E_{01}''(z-z_\iota)^4dz + \frac{15}{32\phi_0^4}\int E_{01}^3E_{02}(z-z_\iota)^4dz.
\end{aligned}
$$

The coefficients A_{abcd} of eq. (26) are obtained by replacing $E_{01}(z)$ by $E_{02}(z)$, a by b, c by d, and vice-versa.

The coefficients of the expressions for Δx_ι and Δy_ι for a single electrostatic deflection field may be obtained by assuming one of the field-strength distributions $E_{01}(z)$ or $E_{02}(z)$ to be identically equal to zero.

If, for example, $E_{02}(z)$ is identically equal to zero, the only terms remaining are those in the brackets { } of eq. (27). These terms are called β_{abcd} and they are identical with those published by Hutter.[5] The terms of the coefficients A_{abcd} which remain if $E_{02}(z) \equiv 0$ are called α_{abcd}.

The expression given by eq. (24) may be used to determine the "deflection" if the actual field-strength distributions $E_{01}(z)$ and $E_{02}(z)$ are known. For large distances from the deflection system, the expression for the deflection in one direction is equivalent to a formula derived by Rüdenberg.[6] For a deflection in the y direction this expression is

$$d = Y_i = (z_i - z_2) \cdot \frac{c\Delta\phi}{2\phi_0\epsilon_0}, \tag{28}$$

where $\rho = c\Delta\phi$ is the charge density, c is the capacitance per unit width, and ϵ_0 is the dielectric constant of free space, z_i is the position coordinate of the screen, z_2 is an axial coordinate where $E(z)$ is essentially equal to zero. The field-strength distribution $E(z)$ or the capacitance c can only be determined, in most practical cases, by experiment. An electrolytic tank potential-plotting device for the determination of $E(z)$ was described by Hutter[7] and graphs for the deflection were prepared for a number of different deflection fields such as (1) parallel plates with fringing field, (2) parallel cylinders or wires, (3) semi-infinite coplanar sheets, and (4) bent plates.

The expressions for Δx_i and Δy_i may be used to compute spot distortion as well as pattern distortion. The former may be obtained by taking x_{iu}, y_{iu}, x_{iu}', and y_{iu}' as the coordinates and slopes of an electron on the circumference of the undeflected beam at the screen. If this procedure is repeated for several points on the circumference of the fluorescent spot, the shape of the distorted spot may be obtained. Magnitude of pattern distortion may be obtained if these quantities x_{iu}, y_{iu}, x_{iu}' and y_{iu}' are taken as the coordinates and slopes of an infinitely thin, undeflected electron beam at the screen. The results of a computation of spot and pattern distortions for the two types of deflection electrode systems shown in Fig. 5 are shown in Fig. 6. The electron beam diameter was assumed to be .08″ at the point where the beam entered the first deflection field. The beam was taken as conical in shape with the apex of the cone at the center of the screen.

The pattern distortion is shown in the lower left hand corner of Fig. 6. Since the distortions in all four quadrants of the target are the same, the total pattern distortion is seen to be of the pincushion type. The spot distortions at the three points A, B, and C (see areas bounded by broken

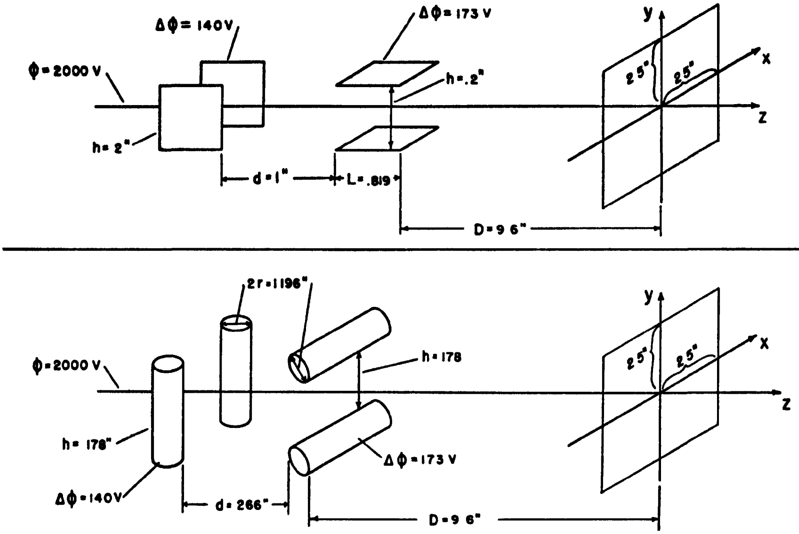


Fig. 5.—The geometry of two crossed, two-dimensional electric deflection systems. The dimensions and potentials are chosen to give equal deflection sensitivity for both systems, one consisting of parallel plates and the other consisting of parallel cylinders. The distortion and defocusing effects of these two systems are compared and the result of the comparison shown in Fig. 6.

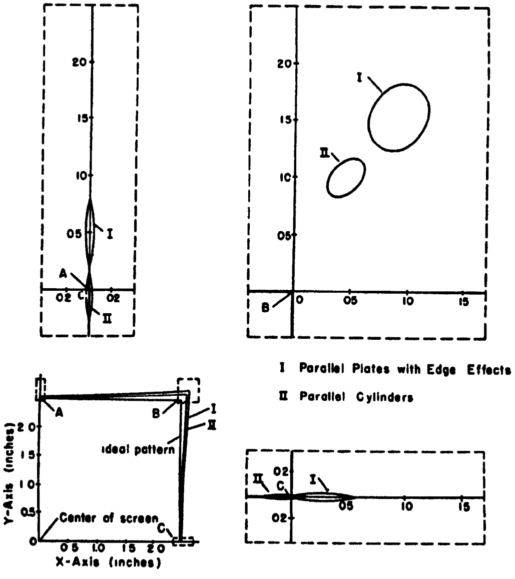


Fig. 6.—Pattern and spot distortions produced by two types of deflection fields shown in Fig. 5.[7] (*Courtesy of the Journal of Applied Physics.*)

lines in lower left hand diagram) are shown enlarged at adjacent positions in the figure. It is seen that the spot is elongated in the direction parallel to the direction of deflection due to the cylindrical lens associated with the deflection system, and that the distortions are more pronounced for the vertical deflections resulting from the lower deflection sensitivity. It should also be noted that the distortions in the corner, B, cannot be obtained by any additive process of the distortions produced by each field alone. This is due to the fact that the deflection by the first system gives the beam unfavorable initial conditions when entering the second system.

4. Single Three-Dimensional Electric Deflection Field

The individual deflection fields of crossed systems discussed so far were assumed to be independent of one coordinate, i.e., so-called two-dimensional fields. We shall now discuss a field of the type described under (5).

The path differential equations in this case are

$$\left.\begin{aligned}
&\frac{d}{dz}\left[x' - \frac{1}{2}\frac{E_0}{\phi_0}yx' - \frac{1}{8}\frac{E_0^2}{\phi_0^2}y^2x' - \frac{1}{2}(x'^2 + y'^2)x'\right] - \\
&\qquad\qquad - \left[-\frac{E_2}{\phi_0}xy\right] = 0, \\
&\frac{d}{dz}\left[y' - \frac{1}{2}\frac{E_0}{\phi_0}yy' - \frac{1}{8}\frac{E_0^2}{\phi_0^2}y^2y' - \frac{1}{2}(x'^2 + y'^2)y'\right] - \\
&\quad - \left[-\frac{1}{2}\frac{E_0}{\phi_0} - \frac{1}{4}\frac{E_0^2}{\phi_0^2}y + \frac{1}{4\phi_0}(E_0'' + 2E_2)y^2 - \right. \\
&\quad - \frac{1}{2}\frac{E_0}{\phi_0}x^2 - \frac{3}{16}\frac{E_0^3}{\phi_0^3}y^2 - \frac{1}{4}\frac{E_0}{\phi_0}(x'^2 + y'^2) - \\
&\qquad\qquad \left. - \frac{1}{8}\frac{E_0^2}{\phi_0^2}y(x'^2 + y'^2)\right] = 0.
\end{aligned}\right\} \tag{29}$$

If the assumption is now made that the field varies gradually in the x direction, i.e.,

$$\left|\frac{E_2}{E_0}x^2\right| \ll 1, \tag{30}$$

one gets as a first step:

$$x'' = 0, \qquad y'' = -\frac{1}{2}\frac{E_0}{\phi_0} \tag{31}$$

for small angle deflections, i.e., if one also assumes

$$\left|\frac{1}{2}\frac{E_0}{\phi_0}y\right| \ll 1, \tag{32}$$

the solutions of eq. (31) are

$$\left.\begin{aligned} x_g(z) &= x_{iu} + x_{iu}'(z - z_i), \\ y_g(z) &= y_{iu} + y_{iu}'(z - z_i) + Y(z), \end{aligned}\right\} \tag{33}$$

where

$$Y(z) = -\frac{1}{2\phi_0}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} E_0(u)du. \tag{33a}$$

The differential equations permitting the determination of higher-order terms are:

$$\left.\begin{aligned} x'' &= \frac{d}{dz}\left[\frac{1}{2}\frac{E_0}{\phi_0}y_gx_g' + \frac{1}{8}\frac{E_0^2}{\phi_0^2}y^2_gx_g' + \frac{1}{2}(x_g'^2 + y_g'^2)x_g'\right] - \frac{E_2}{\phi_0}x_gy_g, \\ y'' &= \frac{d}{dz}\left[\frac{1}{2}\frac{E_0}{\phi_0}y_gy_g' + \frac{1}{8}\frac{E_0^2}{\phi_0^2}y_g^2y_g' + \frac{1}{2}(x_g'^2 + y_g'^2)y_g'\right] - \\ &\quad - \frac{1}{2}\frac{E_0}{\phi_0} - \frac{1}{4}\frac{E_0^2}{\phi_0^2}y_g + \frac{1}{4\phi_0}(E_0'' + 2E_2)y_g^2 - \frac{1}{2}\frac{E_2}{\phi_0}x_g^2 - \\ &\quad - \frac{3}{16}\frac{E_0^3}{\phi_0^3}y_g^2 - \frac{1}{4}\frac{E_0}{\phi_0}(x_g'^2 + y_g'^2) - \frac{E_0^2}{8\phi_0^2}y_g(x_g'^2 + y_g'^2). \end{aligned}\right\} \tag{34}$$

The additional deflection terms Δx_i and Δy_i may then be determined. The coefficients have the following form:

$$\left.\begin{aligned} \alpha_{0000}^{(n)} &= 0 + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2Y(z - z_i)dz, \\ \alpha_{0010}^{(n)} &= \alpha_{0010} + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2Y(z - z_i)^2dz, \\ \alpha_{1100}^{(n)} &= 0 + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2Y(z - z_i)dz, \\ \alpha_{0110}^{(n)} &= \alpha_{0110} + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^2dz, \\ \alpha_{1001}^{(n)} &= 0 + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^2dz, \\ \alpha_{0011}^{(n)} &= \alpha_{0011} + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^3dz, \\ \alpha_{0210}^{(n)} &= \alpha_{0210}, \qquad \alpha_{0111}^{(n)} = \alpha_{0111}, \qquad \alpha_{0012}^{(n)} = \alpha_{0012}, \end{aligned}\right\} \tag{35}$$

$$\left.\begin{aligned}
\beta_{0000}{}^{(n)} &= \beta_{0000} - \frac{1}{2\phi_0}\int_{z_0}^{z_i} E_2Y^2(z - z_i)dz,\\
\beta_{0100}{}^{(n)} &= \beta_{0100} - \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2Y(z - z_i)dz,\\
\beta_{0001}{}^{(n)} &= \beta_{0001} - \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2Y(z - z_i)^2dz,\\
\beta_{2000}{}^{(n)} &= 0 + \frac{1}{2\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)dz,\\
\beta_{0200}{}^{(n)} &= \beta_{0200} - \frac{1}{2\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)dz,\\
\beta_{0020}{}^{(n)} &= \beta_{0020} + \frac{1}{2\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^3dz,\\
\beta_{0002}{}^{(n)} &= \beta_{0002} - \frac{1}{2\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^3dz,\\
\beta_{1010}{}^{(n)} &= 0 + \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^2dz,\\
\beta_{0101}{}^{(n)} &= \beta_{0101} - \frac{1}{\phi_0}\int_{z_0}^{z_i} E_2(z - z_i)^2dz,\\
\beta_{0300}{}^{(n)} &= \beta_{0300}, \qquad \beta_{0120}{}^{(n)} = \beta_{0120}, \qquad \beta_{0102}{}^{(n)} = \beta_{0102},\\
\beta_{0021}{}^{(n)} &= \beta_{0021}, \qquad \beta_{0201}{}^{(n)} = \beta_{0201}, \qquad \beta_{0003}{}^{(n)} = \beta_{0003},
\end{aligned}\right\} \tag{36}$$

where the quantities α_{abcd} and β_{abcd} are as defined before.

As an example, let us examine the magnitudes of pattern and spot distortion of a deflection field having the field-strength distribution in the xy-plane given by

$$E_y|_{y=0} = -E_m \frac{1}{1 + \left(\frac{z}{a}\right)^2 + \left(\frac{x}{b}\right)^2}. \tag{37}$$

Expanding $E_y(x, z)$ in a power series of powers of x, we get

$$E_y|_{y=0} = -E_m\left\{\frac{1}{1 + \left(\frac{z}{a}\right)^2} - \frac{1}{b^2}\frac{1}{\left[1 + \left(\frac{z}{a}\right)^2\right]^2}x^2 + - \quad \cdot\cdot\right\}. \tag{38}$$

We have then:

$$\left.\begin{aligned}
E_0(z) &= -\frac{E_m}{1 + \left(\frac{z}{a}\right)^2},\\
E_2(z) &= \frac{1}{b^2}\frac{E_m}{\left[1 + \left(\frac{z}{a}\right)^2\right]^2}.
\end{aligned}\right\} \tag{39}$$

One may now choose a pair of typical values for a and E_m:

$$a = 0.34, \qquad E_m = 939,$$

Since $y_{iu} = y_{iu}' = 0$ and $x_{iu}' \neq 0$, $x_{iu} \neq 0$ for the second deflection field, it follows that the additional deflections Δx_i and Δy_i are given by

$$\left.\begin{aligned} \Delta x_i &= \alpha_{1000}{}^{(n)}x_{iu} + \alpha_{0010}{}^{(n)}x_{iu}', \\ \Delta y_i &= \beta_{0000}{}^{(n)} + \beta_{2000}{}^{(n)}x_{iu}{}^2 + \beta_{0020}{}^{(n)}x_{iu}'^2 + \beta_{1010}{}^{(n)}x_{iu}x_{iu}. \end{aligned}\right\} \tag{40}$$

Choosing $x_{iu} = 2.5''$, $x_{iu}' = 0.20$ we obtain for Δx_i and Δy_i,

$$\Delta x_i = 0.0502 - 0.126\frac{1}{b^2}, \qquad \Delta y_i = 0.09224 - 0.255\frac{1}{b^2}.$$

For $b = 1.58$, the values of Δx_i and Δy_i are nearly equal to zero, hence the pattern distortion is considerably reduced. The magnitude of spot distortion is nearly the same for the two cases $b = 1.58$ and $b = \infty$.

5. *Crossed Unbalanced Electric Deflection Fields*

The distortion and defocusing effects of an unbalanced field are larger than those produced by a balanced field. Two such crossed fields produce a characteristic pattern distortion, the pattern having the shape of a diamond. The general theoretical method used throughout this paper may be applied to such a case to give an explanation for this type of distortion. Higher-order effects in the pattern distortion and the spot distortion will not be discussed here.

The potential distribution is described by eqs. (13) and (14). The first-order differential equations for an electron path are then given by

$$\left.\begin{aligned} &\frac{d}{dz}\left[x'\left\{1 + \frac{1}{2\phi_m}(\phi_{m1}\phi_1(z) + \phi_{m2}\phi_2(z))\right\}\right] = \\ &\qquad = -\frac{1}{2\phi_m}E_{02}(z)\left\{1 - \frac{1}{2\phi_m}(\phi_{m1}\phi_1(z) + \phi_{m2}\phi_2(z))\right\}, \\ &\frac{d}{dz}\left[y'\left\{1 + \frac{1}{2\phi_m}(\phi_{m1}\phi_1(z) + \phi_{m2}\phi_2(z))\right\}\right] = \\ &\qquad = -\frac{1}{2\phi_m}E_{01}(z)\left\{1 - \frac{1}{2\phi_m}(\phi_{m1}\phi_1(z) + \phi_{m2}\phi_2(z))\right\}. \end{aligned}\right\} \tag{41}$$

If $x_g = y_g = x_g' = y_g' = 0$ at $z = z_0$, the solutions may be written

$$\left.\begin{aligned} x_g(z) = &-\frac{1}{2\phi_m}\int_{z_0}^{z}\left[1-\frac{1}{2\phi_m}(\phi_{m1}\phi_1(\xi)+\phi_{m2}\phi_2(\xi))\right]d\xi\int_{z_0}^{\xi}E_{02}(u) \\ &\qquad\left[1-\frac{1}{2\phi_m}(\phi_{m1}\phi_1(u)+\phi_{m2}\phi_2(u))\right]du, \\ y_g(z) = &-\frac{1}{2\phi_m}\int_{z_0}^{z}\left[1-\frac{1}{2\phi_m}(\phi_{m1}\phi_1(\xi)+\phi_{m2}\phi_2(\xi))\right]d\xi\int_{z_0}^{\xi}E_{01}(u) \\ &\qquad\left[1-\frac{1}{2\phi_m}(\phi_{m1}\phi_1(u)+\phi_{m2}\phi_2(u))\right]du, \end{aligned}\right\} \qquad (42)$$

or, ignoring terms of $\phi_{m\nu}/\phi_m$ which are of higher order than the first,

$$\left.\begin{aligned} x_g(z) = &-\frac{1}{2\phi_m}\int_{z_0}^{z}d\xi\int_{z_0}^{\xi}E_{02}(u)du+\frac{1}{2\phi_m}\int_{z_0}^{z}d\xi\int_{z_0}^{\xi}\frac{1}{2\phi_m}(\phi_{m1}\phi_1+ \\ &+\phi_{m2}\phi_2)E_{02}du+\frac{1}{2\phi_m}\int_{z_0}^{z}\frac{1}{2\phi_m}(\phi_{m1}\phi_1+ \\ &\qquad\qquad+\phi_{m2}\phi_2)d\xi\int_{z_0}^{\xi}E_{02}du, \\ y_g(z) = &-\frac{1}{2\phi_m}\int_{z_0}^{z}d\xi\int_{z_0}^{\xi}E_{01}(u)du+\frac{1}{2\phi_m}\int_{z_0}^{z}d\xi\int_{z_0}^{\xi}\frac{1}{2\phi_m}(\phi_{m1}\phi_1+ \\ &+\phi_{m2}\phi_2)E_{01}du+\frac{1}{2\phi_m}\int_{z_0}^{z}\frac{1}{2\phi_m}(\phi_{m1}\phi_1+ \\ &\qquad\qquad+\phi_{m2}\phi_2)d\xi\int_{z_0}^{\xi}E_{01}du. \end{aligned}\right\} \qquad (43)$$

Defining

$$E_{01}(z) = \Delta\phi_1 f_1(z), \qquad E_{02}(z) = \Delta\phi_2 f_2(z), \qquad (44)$$

we obtain for the deviations from a rectangular pattern:

$$\left.\begin{aligned} \Delta x_i &= \Delta\phi_2\phi_{m1}\cdot A_{12}+\Delta\phi_2\phi_{m2}A_{22}, \\ \Delta y_i &= \Delta\phi_1\phi_{m1}\cdot A_{11}+\Delta\phi_1\phi_{m2}A_{21}, \end{aligned}\right\} \qquad (45)$$

where

$$A_{\kappa\lambda} = \frac{1}{4\phi_m^2}\left\{\int_{z_0}^{z_i}d\xi\int_{z_0}^{\xi}\phi_\kappa(u)f_\lambda(u)du+\int_{z_0}^{z_i}\phi_\kappa(\xi)d\xi\int_{z_0}^{\xi}f_\lambda(u)du.\right. \qquad (46)$$

Fig. 7 shows the ideal pattern (heavy line) and the distorted pattern (broken line) for the case of two crossed but otherwise identical electric unbalanced deflection fields. In this case $f_1 = f_2$, $|\Delta\phi_1| = [\Delta\phi_2|$, $\phi_1 = \phi_2$, and $|\phi_{m1}| = |\phi_{m2}|$. Unlike the pattern distortion due to balanced fields, the distortion here is seen to be of a different character in each quadrant of the target. If the two crossed fields are not identical, e.g., if they are displaced along the optical axis with respect to each other, the symmetry of the pattern about the line $y = x$ will disappear.

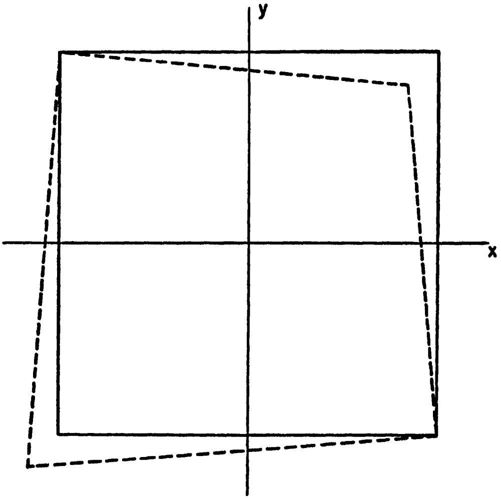


FIG. 7.—Pattern distortion, produced by two crossed unbalanced electric deflection fields.

6. *Crossed Magnetic Deflection Fields*

The magnetic field strength distribution is given by eq. (16). The deflection path of an electron through any magnetic field was given by eq. (18). They may also be written in the form

$$\left.\begin{aligned}\sqrt{\phi}\,\frac{d}{dz}\,[x'/\sqrt{1 + x'^2 + y'^2}] &= \eta[-y'H_z + H_y],\\ \sqrt{\phi}\,\frac{d}{dz}\,[y'/\sqrt{1 + x'^2 + y'^2}] &= \eta[x'H_z - H_x],\end{aligned}\right\} \qquad (47)$$

where ϕ is the potential of the space in which the electrons are moving and $\eta = \sqrt{e/2m}$, e and m being respectively the charge and mass of the electron. For a magnetic field given by eq. (16), we obtain

$$\left.\begin{aligned}\frac{d}{dz}\left[x' - \frac{1}{2}x'^3 - \frac{1}{2}x'y'^2\right] &= \frac{\eta}{\sqrt{\phi}}\Big[H_{02}'y'x - H_{01}'y'y + H_{01} + \\ &\qquad + H_{22}x^2 - 2H_{22}xy - \left(H_{21} + \frac{1}{2}H_{01}''\right)y^2\Big], \\ \frac{d}{dz}\left[y' - \frac{1}{2}y'^3 - \frac{1}{2}x'^2y'\right] &= \frac{\eta}{\sqrt{\phi}}\Big[-H_{02}'xx' + H_{01}'x'y + H_{02} - \\ &\qquad - \left(H_{22} + \frac{1}{2}H_{02}''\right)x^2 - 2H_{22}xy + H_{22}y^2\Big],\end{aligned}\right\} \tag{48}$$

where only terms of zero, first, and second order of x, x', y and y', are considered. To a first approximation, the equations of motion are given by

$$\left.\begin{aligned}x'' &= \frac{\eta}{\sqrt{\phi}}H_{01}, \\ y'' &= \frac{\eta}{\sqrt{\phi}}H_{02}.\end{aligned}\right\} \tag{49}$$

Solution of these equations yields

$$\left.\begin{aligned}x_g(z) &= x_{iu} + x_{iu}'(z - z_i) + X(z), \\ y_g(z) &= y_{iu} + y_{iu}'(z - z_i) + Y(z),\end{aligned}\right\} \tag{50}$$

where

$$\left.\begin{aligned}X(z) &= \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} H_{01}(u)du, \\ Y(z) &= \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} H_{02}(u)du.\end{aligned}\right\} \tag{51}$$

The quantities $X = X(z)$ and $Y = Y(z)$ at $z = z_s$ (the screen position) are the deflections at the screen. The deflections in the two directions x and y are independent of each other in this first-order approximation, since one depends only upon H_{01} the other only upon H_{02}.

Following the procedure used for electric deflection fields, x, x', y and y' of eq. (48) will be replaced by x_g, x_g', y_g and y_g' in all terms of higher order than the first. The solutions of the resulting differential equations can be obtained by simple integrations. If the differences $\Delta^M x_i$ and $\Delta^M y_i$ between the new solutions and those given by eq. (50) are taken, one obtains

$$\left.\begin{aligned}\Delta^M x_i &= \sum_{abcd}^{0\ldots2} \gamma_{abcd} x_{iu}^{\,a} y_{iu}^{\,b} x_{iu}'^{\,c} y_{iu}'^{\,d}, \\ \Delta^M y_i &= \sum_{abcd}^{0\ldots2} \delta_{abcd} x_{iu}^{\,a} y_{iu}^{\,b} x_{iu}'^{\,c} y_{iu}'^{\,d}.\end{aligned}\right\} \tag{52}$$

The coefficients γ_{abcd} of eq. (52) are

$$\left.\begin{aligned}
\gamma_{0000} = {} & \frac{1}{2}\int_{z_0}^{z_i} X'^3 dz + \frac{1}{2}\int_{z_0}^{z_i} X'Y'^2 dz - \\
& - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y'XH_{02}'(z-z_i)dz + \\
& + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YY'H_{01}'(z-z_i)dz - \\
& - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} H_{21}X^2(z-z_i)dz + \\
& + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y^2H_{21}(z-z_i)dz + \\
& + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XYH_{22}(z-z_i)dz + \\
& \qquad\qquad + \frac{1}{2}\frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y^2H_{01}''(z-z_i)dz, \\
\gamma_{1000} = {} & - \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XH_{21}(z-z_i)dz + \\
& + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{22}(z-z_i)dz - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y'H_{02}'(z-z_i)dz, \\
\gamma_{0100} = {} & \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y'H_{01}'(z-z_i)dz + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XH_{22}(z-z_i)dz + \\
& + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{21}(z-z_i)dz + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{01}''(z-z_i)dz, \\
\gamma_{0010} = {} & \frac{3}{2}\int_{z_0}^{z_i} X'^2 dz + \frac{1}{2}\int_{z_0}^{z_i} Y'^2 dz - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y'H_{02}'(z-z_i)^2 dz - \\
& - \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XH_{21}(z-z_i)dz + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{22}(z-z_i)^2 dz, \\
\gamma_{0001} = {} & \int_{z_0}^{z_i} X'Y'dz - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XH_{02}'(z-z_i)dz + \\
& + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{01}'(z-z_i)dz + \\
& + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} Y'H_{01}'(z-z_i)^2 dz + \\
& + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} XH_{22}(z-z_i)^2 dz + \frac{2\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{21}(z-z_i)^2 dz + \\
& \qquad\qquad + \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} YH_{01}''(z-z_i)^2 dz, \\
\gamma_{2000} = {} & - \frac{\eta}{\sqrt{\phi}}\int_{z_0}^{z_i} H_{21}(z-z_i)dz,
\end{aligned}\right\} \qquad (53)$$

$$\left.\begin{aligned}
\gamma_{1100} &= \frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{22}(z - z_i)dz,\\
\gamma_{0200} &= \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{21}(z - z_i)dz + \frac{1}{2}\frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{01}''(z - z_i)dz,\\
\gamma_{0020} &= \frac{3}{2}\int_{z0}^{z_i} X'dz - \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{21}(z - z_i)^3dz,\\
\gamma_{0011} &= \int_{z0}^{z_i} Y'dz + \frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{22}(z - z_i)^3dz -\\
&\qquad - \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{02}'(z - z_i)^2dz,\\
\gamma_{0110} &= \frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{22}(z - z_i)^2dz,\\
\gamma_{1001} &= -\frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{02}'(z - z_i)dz + \frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{22}(z - z_i)^2dz,\\
\gamma_{1010} &= -\frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{21}(z - z_i)^2dz,\\
\gamma_{0101} &= \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{01}'(z - z_i)dz + \frac{2\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{21}(z - z_i)^2dz +\\
&\qquad + \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{01}''(z - z_i)^2dz,\\
\gamma_{0002} &= \frac{1}{2}\int_{z0}^{z_i} X'dz + \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{01}'(z - z_i)^2dz +\\
&\quad + \frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{21}(z - z_i)^3dz + \frac{1}{2}\frac{\eta}{\sqrt{\phi}}\int_{z0}^{z_i} H_{01}''(z - z_i)^3dz.
\end{aligned}\right\} \quad (53)$$

The coefficients δ_{abcd} of eq. (52) are obtained by replacing $H_{2n,1}$ by $H_{2n,2}$, a by b, c by d, and vice-versa. Expressions for Δx_i and Δy_i in the case of a single magnetic field may be obtained by assuming one of the two field-strength distributions to be equal to zero.

Eq. (51) may be used to compute the deflection if the axial field-strength distributions for both deflection fields are known. To illustrate spot and pattern distortions produced by magnetic fields, we take the following field functions:

$$H_x = -\frac{H_m}{1 + \left(\frac{z}{b}\right)^2 + \left(\frac{y}{a}\right)^2}, \qquad H_y = \frac{H_m}{1 + \left(\frac{z}{b}\right)^2 + \left(\frac{x}{a}\right)^2}; \tag{54}$$

it may be verified that

$$H_{01} = \frac{H_m}{\left[1 + \left(\frac{z}{b}\right)^2\right]}, \qquad H_{21} = -\frac{1}{a^2}\frac{H_m}{\left[1 + \left(\frac{z}{b}\right)^2\right]^2}, \tag{55}$$

and

$$H_{02} = \frac{H_m}{\left[1 + \left(\frac{z}{b}\right)^2\right]}, \qquad H_{22} = -\frac{1}{a^2}\frac{H_m}{\left[1 + \left(\frac{z}{b}\right)^2\right]^2}. \tag{56}$$

The distortions[7] shown in Fig. 8 were calculated for the following set of parameters:

$$\phi = 2000V, \qquad H_m = 35.4 \text{ gauss}$$

$$b = 0.34'', \qquad z_i = 10.198'', \qquad \frac{z_i}{b} = 30.$$

and two values of a, namely, $a_1 = 1.36''$, $a_2 = 0.68''$. The maximum deflections on the screen were $d_V = 2.5''$, $d_h = 2.5''$. The undeflected electron beam was assumed to be cylindrical with radius $r_b = 0.04''$.

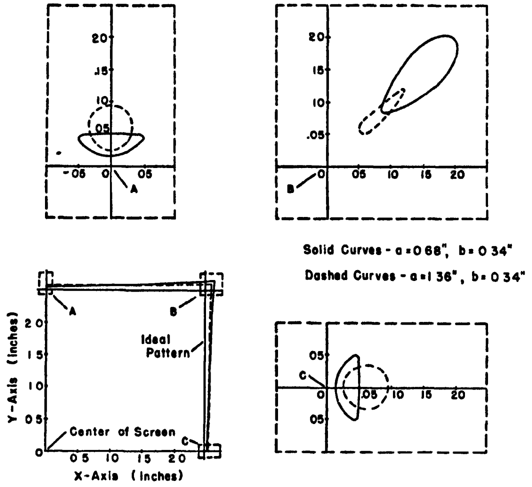


FIG. 8.—Pattern and spot distortion by a special pair of crossed magnetic fields. The fields are given by:[7]

$$H_x = -\frac{H_m}{1 + \left(\frac{z}{b}\right)^2 + \left(\frac{y}{a}\right)^2}, \qquad H_y = \frac{H_m}{1 + \left(\frac{z}{b}\right)^2 + \left(\frac{x}{a}\right)^2}.$$

The notation is explained in the text. (*Courtesy of the Journal of Applied Physics.*)

The pattern distortion is shown for one quadrant of the target in the lower left hand corner of Fig. 8. The distortions of the spot, circular in shape at the center of the target, at A, B, and C, are drawn on a larger scale in the adjacent diagrams. It can be seen that the spot is compressed in the direction of deflection if the beam is deflected by a single field but that it is elongated when deflected by both fields simultaneously. A qualitative explanation of the focusing action of a single magnetic deflection field is based on the fact that for vertically upward deflection the lower beam edge passes through regions of higher field strength than does the upper edge. The under focusing or diverging action of two crossed fields of the type given by eq. (54) is due to the fact that the field strength distribution along the line $y = x$ at $z = 0$ is saddle shaped. Hence the inner edge of the beam passes through regions of weaker fields than does the outer edge.

It should furthermore be noted that spot as well as pattern distortions increase in magnitude if the half-width parameter "a" is increased.

7. Correction of Spot and Pattern Distortion

It has been mentioned that unbalanced electrostatic deflection fields cause a trapeozidal pattern distortion. A number of suggestions have been made as to means for reducing this effect. With other deflection fields the pattern distortion is considered to be less serious than the spot distortion, hence more effort has been directed to a reduction of the latter.

We have seen that the coefficients given by the eqs. (27), (35), (36), and (53) each determine a characteristic type of distortion. A method to obtain improved deflection fields might be as follows. The coefficients are computed for a number of different fields and a comparison made to show which is best as far as spot distortion is concerned. The field distribution of the best one of these fields is then modified in such a way as to reduce the values of the coefficients. This procedure may be repeated several times until a set of coefficients having desirably low values is found. The new field-strength function obtained in such a manner then determines the proper configurations of the deflection electrodes or magnets. This method, however, is extremely tedious and has never been applied so far as the author knows.

Another general method may be based on the theory stated by Moss[8] which "is of fundamental importance in the design of all electrostatic cathode ray tubes, since it indicates the general method by which the beam width, and, therefore, the spot density can be increased without an increase in deflection defocusing. Large deflectors and a large neck diameter to accommodate them are used." The theorem reads:

"If the beam width and scale of the whole deflectors (including their spacing) are multiplied by κ, then the increase of spot size on deflection through a constant angle is unchanged, provided the distance between the screen and the center of deflection is also unchanged."

The proof of this theorem was based on theories of scale, energy, and dimensional homogeneity, and upon experimental evidence. A more rigorous proof showing, at the same time, some limitations of the scaling process can be based on the theory developed here.

Consider a point-focused (conical) electron beam. The defocusing effects for a single electrostatic deflection field will be given to a first approximation by

$$\Delta y_i = \beta_{0000} + \beta_{0001} y_{iu}'$$

If all dimensions are now increased by a factor κ (keeping all potentials constant), the deflection d will very nearly increase by this factor κ as long as the electrode-to-screen distance is sufficiently large. The coefficients β_{0000} and β_{0001} will also increase by the same factor and, since y_{iu}' remains unchanged, Δy_i will increase to $\kappa \cdot \Delta y_i$. If the screen is now brought back to its original position and the electron beam is refocused to a point at the center of the screen, the deflection will again be equal to d. The angle y_{iu}', however, will be κ times its original value (for small angles). An inspection of the coefficients β_{0000} and β_{0001} shows that for large electrode-to-screen distances the former is a linear function while the latter is a quadratic function of this distance. Decreasing the electrode-to-screen distance restores, therefore, β_{0000} to its original value. Since β_{0001} increased to κ times its original value, and since it is a quadratic function of the electrode-to-screen distance, $\beta_{0001} y_{iu}'$ will reduce to its original value when the distance is decreased to its original value. Hence Δy_i has the original value.

Space limitation in actual cathode ray and television tubes precludes the possibility of making full use of the two general methods described thus far. Cathode ray tube engineers appear to have become resigned to the fact that deflection systems produce spot distortion and that the modification of the electrode shapes or the magnetic coils cannot bring much improvement. A number of attempts have been made to influence the electron beam in such a way that the beam distortions are compensated.

Two methods will now be described which may be of interest. One of these methods, disclosed by Schlesinger[9,10] in a series of patents, aims at accomplishing a reduction of spot distortion by applying an auxiliary potential difference between the last anode and the electrostatic deflection system, and by changing the potential of the focusing element of the electron gun. These potentials are nonlinear functions of the deflection potential. The potential which is applied to the focusing electrode reduces the refracting power of the electron gun. The potential difference between the deflecting system and the last anode is applied in such a way that the deflection system becomes unbalanced but always so that the momentarily negative plate is closer to the potential of the last anode. Due to this potential difference, a lens is created in the space between the last anode and the deflection system. The first portion of this lens, in the neighborhood of the deflection plates is a two-dimensional diverging lens. If now, at the same time, the refracting power of the focusing system is reduced, we obtain as the total effect of these potentials the distortion of the conical electron beam such that there is an elongation in the direction of the deflection as shown in Fig. 9. This elongation of the spot will then be reduced to zero by the two-dimensional focusing action of the deflection field. Fig. 9 also shows the light-optical arrangement which is the equivalent of the electron-optical arrangement just described. Schlesinger[9,10] has suggested several circuits for deriving proper auxiliary potentials from the deflection voltages.

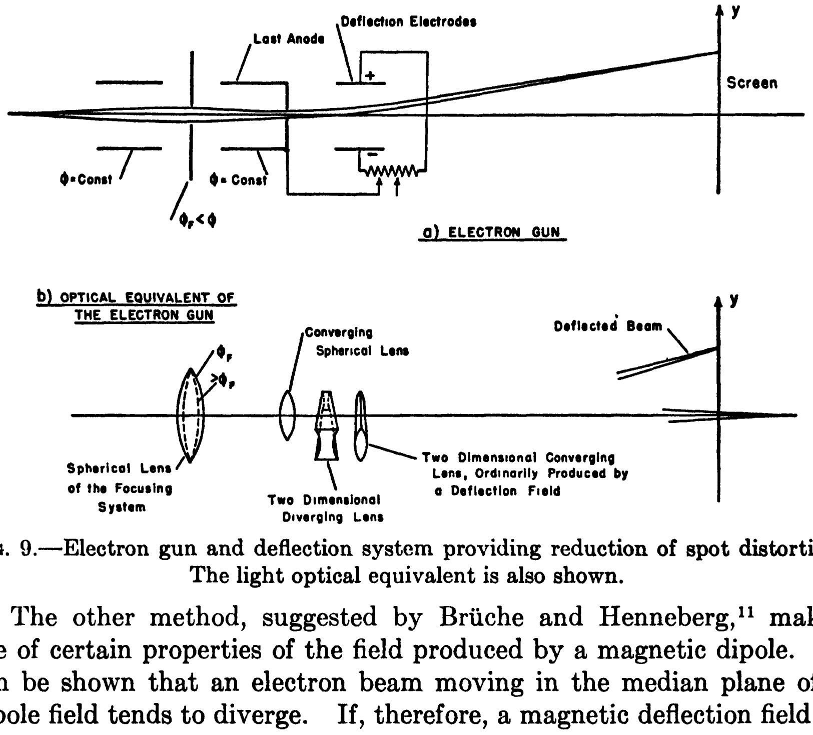

Deflection Electrodes
Last Anode
+
−
y
Screen
Φ=Const
Φ= Const
$\Phi_r < \Phi$
a) ELECTRON GUN
b) OPTICAL EQUIVALENT OF THE ELECTRON GUN
Φ_r
$>\Phi_r$
Converging Spherical Lens
Deflected Beam
y
Spherical Lens of the Focusing System
Two Dimensional Diverging Lens
Two Dimensional Converging Lens, Ordinarily Produced by a Deflection Field

FIG. 9.—Electron gun and deflection system providing reduction of spot distortion. The light optical equivalent is also shown.

The other method, suggested by Brüche and Henneberg,[11] makes use of certain properties of the field produced by a magnetic dipole. It can be shown that an electron beam moving in the median plane of a dipole field tends to diverge. If, therefore, a magnetic deflection field or an electrostatic deflection field is combined with such a dipole field as shown in Figs. 10 and 11 it is possible to adjust the dipole field in relation to any particular deflection field in such a way that the focusing and defocusing properties of the fields compensate each other.

It has been stated that pattern distortion does not represent a serious problem in balanced deflection fields. A method of reducing pattern distortion of unbalanced electrostatic deflection systems was suggested by Fleming-Williams.[12] If the vertical deflection system is a balanced electrostatic or a magnetic one and the horizontal deflection system is an

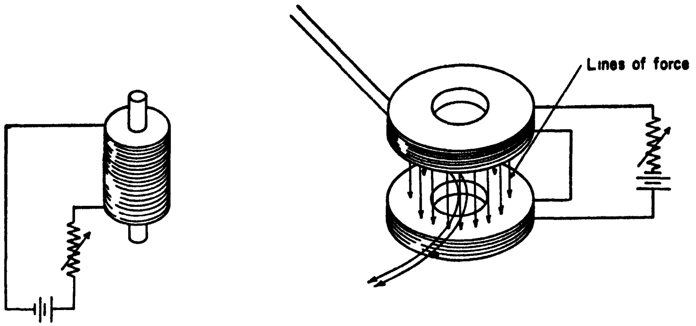


FIG. 10.—Dipole field and magnetic deflection system. An electron beam moving in the equatorial plane of the dipole field tends to diverge, counteracting the converging action of the magnetic deflection field.

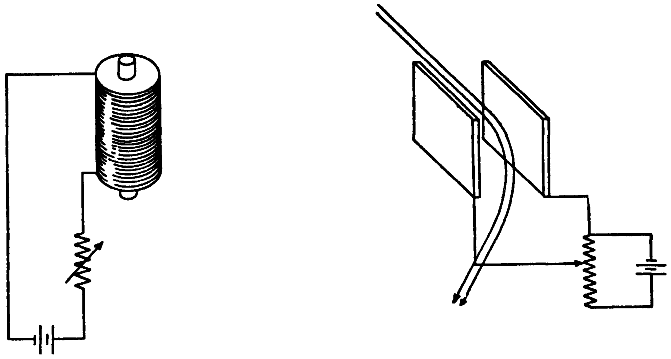
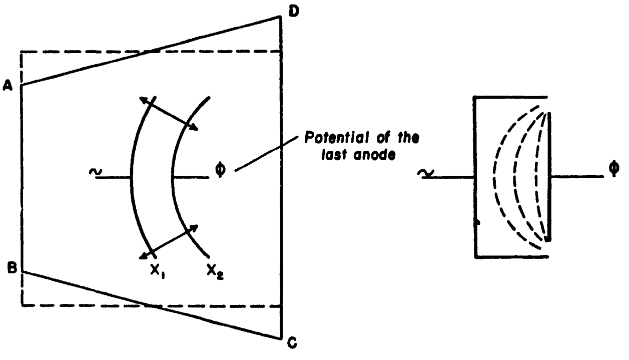


FIG. 11.—Dipole field and electric deflection system. Here the diverging action of the dipole field is used to counteract the converging action of the electric deflection field.



a b

FIG. 12.—Deflection electrodes, designed to yield reduced trapezoidal distortion. (Courtesy of the *Wireless Engineer*, see reference 12.)

unbalanced one, the pattern will be that of a trapezoid *ABCD* shown in Fig. 12a. The horizontal deflection electrodes are indicated in this figure by X_1 and X_2. The electrode X_2 is assumed to be the one connected to the last anode of the electron gun. Curving these electrodes as shown produces an electrostatic field which exerts a focusing action counteracting that of the deflection field. The additional focusing action is indicated by the arrows in Fig. 12a. A similar effect is produced by an electrode system shown in Fig. 12b, as may easily be understood from the equipotential line distribution.

8. *Ion Traps and Linear Mass Spectrometers*

A combination of electrostatic and magnetic deflection fields has been used in instruments designed to separate various kinds of charged particles which are emitted by a source. The small-angle deflection theory does not apply to most of these instruments. In two of them, however, the angle of deflection is kept very small for particles having a specified ratio of charge to mass. They are the ion trap and the linear mass spectrometer.

Ion traps[13,14] are used in television tubes to separate negative ions from electrons in order to prevent the ions from reaching the fluorescent screen where they would be detrimental to the fluorescent properties of the screen material. Ordinarily the total beam is deflected off the axis by a magnetic field and thereafter an electric field is used to return the electron beam to its original direction. Negative ions are not deflected back due to their heavier mass. In most engineering designs of such ion trap arrangements, even where the electric and magnetic fields are established in the same region, two deflections take place and although the angle is kept small, a beam distortion occurs.

The linear mass-cspetrometer[15] employs two superimposed crossed fields, one of which is usually a uniform electric, the other a uniform magnetic field. The directions of the lines of force are such that for a particular kind of ion, characterized by a certain ratio e_1/m_1 of charge to mass, the total deflection is equal to zero. Particles of other e/m ratios also present in the beam are deflected through large angles to either side of the undeflected beam.

The small-angle deflection theory may be applied to the motion of those charged particles for which the total deflection is equal to zero, i.e., electrons in case of ion traps and particles having the ratio e_1/m_1, in the case of linear mass spectrometers. The first-order expression for the deflection of a charged particle of the ratio e_1/m_1 of charge to mass and the speed of ϕ electron volts is given by:

$$Y(z) = -\frac{1}{2\phi}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} E(u)du - \frac{1}{\sqrt{\phi}}\sqrt{\frac{e_1}{2m_1}}\int_{z_0}^{z} d\xi \int_{z_0}^{\xi} H_{02}(u)du. \qquad (57)$$

If both terms of eq. (57) are made numerically equal for every point z of the axis, the charged particle will move in a straight line. A beam of such charged particles will be less distorted than in the cases where a deflection does take place, even if the fields are adjusted for vanishing total deflection. Ion trap fields violated the just mentioned condition as was stated above while the fields of linear mass spectrometers are usually not designed to satisfy this condition in the regions of fringing fields. Oliphant, Shire, and Crowther[15] experienced problems in gun alignment due to the fact that these fringing fields were not designed to compensate each other.

III. LARGE-ANGLE DEFLECTION

It has been shown that deflection-type fields exert also a focusing action on an electron or ion beam. It was shown that the lens action of the deflection field increases with the magnitude of the deflection. This is an undesirable effect in small-angle deflection devices. The purpose of devices using large-angle deflections, however, such as mass spectrometers, make such a focusing action a highly desirable feature. Thus, in the design of cathode ray tubes attempts are made to avoid the focusing action while designs of mass spectrometer fields are governed by the desire to obtain as perfect a focusing action as possible. This means that all ions or electrons leaving either a point source or a line source of certain size should be united again at a point or line after deflection. In case of a line source, one usually does not attempt to design a deflection field yielding a stigmatic imagery. The best that can be hoped for in this case is that a line in the object-space corresponds to a line in the image-space. Point sources are used in case of so-called two directional focusing spectrometers. The fields of such mass spectrometers provide a point-to-point relation between object and image space.

The theory of large-angle deflection fields, as developed by Wendt,[16] is a generalization of the theory of ordinary focusing fields. In case of rotationally symmetrical lens fields the optical axis is a straight line which connects the center of the object plane with the center of the image plane. This axis is at the same time a special electron path. The electromagnetic fields are expanded in series about the optical axis and the equations of motion are integrated by methods of successive approximation. In this manner relations between object and image (such as location and magnification) are obtainable, and mathematical expressions for lens aberration may be derived. In case of large-angle deflection fields, an electron path of certain simple geometrical properties (i.e., circular or cycloidal), is singled out and called the optical axis. The electromagnetic fields are then expanded in a series about this new optical axis and the paths of electrons or ions in the neighborhood of the axis are investigated.

1. Motion of Particles in Systems with Arbitrarily Curved Axes

In order to obtain a general theory, it shall be permitted that the optical axis may be any curve in space. This optical axis is called the w-axis. The two other coordinates called u and v are along lines perpendicular to the space curve at any point. The u-axis is taken in the direction of the principal normal and the v-axis is taken in the direction of the binormal. If ρ is the radius of curvature of the space curve, and τ is the torsion of the optical axis, the line element in the uvw space is given by:

$$dl_w = dw \cdot \sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}. \tag{58}$$

The differential operators in the new system of coordinates are given by:

$$\left.\begin{aligned}
\nabla_u \psi &= \frac{\partial \psi}{\partial u}, \\
\nabla_v \psi &= \frac{\partial \psi}{\partial v}, \\
\nabla_w \psi &= \frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}} \frac{\partial \psi}{\partial w},
\end{aligned}\right\} \tag{59}$$

$$\nabla \cdot \vec{S} = \frac{\begin{aligned}&\frac{\partial}{\partial u}\left[S_u \sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right] + \\ &\qquad + \frac{\partial}{\partial v}\left[S_v \sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right] + \frac{\partial S_w}{\partial w}\end{aligned}}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}}, \tag{60}$$

$$\begin{aligned}
\nabla \cdot \nabla \psi = {} & \frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}} \\
& \left\{\frac{\partial}{\partial u}\left[\frac{\partial \psi}{\partial u}\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right] + \right. \\
& + \frac{\partial}{\partial v}\left[\frac{\partial \psi}{\partial v}\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right] + \\
& \left. + \frac{\partial}{\partial w}\left[\frac{\partial \psi}{\partial w}\frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}}\right]\right\},
\end{aligned} \tag{61}$$

$$\left.\begin{aligned}
[\nabla \times \vec{S}]_u &= \frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}} \\
&\qquad \left\{\frac{\partial}{\partial v}\left[S_w\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right] - \frac{\partial Sv}{\partial w}\right\}, \\
[\nabla \times \vec{S}]_v &= \frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}} \\
&\qquad \left\{\frac{\partial S_u}{\partial w} - \frac{\partial}{\partial u}\left[S_w\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}\right]\right\}, \\
[\nabla \times \vec{S}]_w &= \frac{\partial S_v}{\partial u} - \frac{\partial S_u}{\partial v}.
\end{aligned}\right\} \quad (62)$$

The electric or magnetic potential may then be expanded formally in the following series:

$$\left.\begin{aligned}
\psi &= \psi_{00} + \psi_{10}u + \psi_{01}v + \psi_{20}u^2 + \psi_{11}uv + \psi_{02}v^2 + \cdots, \\
&\text{or} \\
\psi &= \sum_{m,n}^{0\ \infty} \psi_{m,n}(w)\cdot u^m v^n,
\end{aligned}\right\} \quad (63)$$

where the coefficients $\psi_{m,n}$ are functions of w. The components of the field strength are then expanded in the following series:

$$\left.\begin{aligned}
-\frac{\partial\psi}{\partial u} &= -\sum_{m,n}^{0\ \infty} m\psi_{m,n}u^{m-1}v^n, \\
-\frac{\partial\psi}{\partial v} &= -\sum_{m,n}^{0\ \infty} n\psi_{m,n}u^m v^{n-1}, \\
-\frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}}\frac{\partial\psi}{\partial w} &= \\
&= -\frac{1}{\sqrt{\left(1+\frac{u}{\rho}\right)^2 + (u^2+v^2)\tau^2}}\sum_{m,n}^{0\ \infty}\frac{d\psi_{m,n}}{dw}u^m v^n.
\end{aligned}\right\} \quad (64)$$

The magnetic and electrostatic potentials have to satisfy LaPlace's equation:

$$\nabla\cdot\nabla\psi = 0. \quad (65)$$

Substituting the series into this equation, recurrence formulas for the coefficients are obtained. The equations of motion may be derived from Hamilton's principle,

$$\delta\int F\,dw = 0, \tag{66}$$

where

$$\begin{aligned} F = {} & \sqrt{\varphi\left[u'^2 + v'^2 + \left(1 + \frac{u}{\rho}\right)^2 + (u^2 + v^2)\tau^2\right]} - \\ & - \eta\left[A_u u' + A_v v' + A_w\sqrt{\left(1 + \frac{u}{\rho}\right)^2 + (u^2 + v^2)\tau^2}\right], \end{aligned} \tag{67}$$

$$\vec{E} = -\nabla\varphi, \qquad \vec{H} = -\nabla\psi_M, \qquad \vec{H} = \nabla \times \vec{A}, \qquad \eta = \sqrt{\frac{e}{2m}},$$

Expanding F in series of powers of u, v, u' and v' one obtains

$$F = F_0 + F_1 + F_2 + F_3 + \cdots, \tag{68}$$

where:

$$\left.\begin{aligned} F_0 = {} & \sqrt{\phi_{00}}, \\ F_1 = {} & \sqrt{\phi_{00}}\left\{\left[\frac{1}{\rho} + \frac{1}{2}\frac{\phi_{10}}{\phi_{00}} + \frac{\eta H_{v00}}{\sqrt{\phi_{00}}}\right]u + \left[\frac{1}{2}\frac{\phi_{01}}{\phi_{00}} - \frac{\eta H_{u00}}{\sqrt{\phi_{00}}}\right]v\right\} = \\ & \qquad = \sqrt{\phi_{00}}\,\{A_1 u + A_2 v\}, \\ F_2 = {} & \sqrt{\phi_{00}}\left\{\left[\frac{\tau^2}{2} + \frac{1}{2}\frac{\phi_{10}}{\rho\phi_{00}} - \frac{1}{8}\frac{\phi_{10}{}^2}{\phi_{00}{}^2} + \frac{1}{2}\frac{\phi_{20}}{\phi_{00}} + \frac{1}{2}\frac{\eta H_{v00}}{\rho\sqrt{\phi_{00}}} + \right.\right. \\ & + \left.\frac{1}{2}\frac{\eta H_{v10}}{\sqrt{\phi_{00}}}\right]u^2 + \left[\frac{1}{2}\frac{\phi_{01}}{\rho\phi_{00}} - \frac{1}{4}\frac{\phi_{10}\phi_{01}}{\phi_{00}{}^2} + \frac{1}{2}\frac{\phi_{11}}{\phi_{00}} - \frac{\eta H_{u10}}{\sqrt{\phi_{00}}} - \right. \\ & - \left.\frac{\eta H_{w00}{}'}{\sqrt{\phi_{00}}} - \frac{\eta H_{v00}}{\rho\sqrt{\phi_{00}}}\right]uv + \left[\frac{\tau^2}{2} - \frac{1}{8}\frac{\phi_{01}{}^2}{\phi_{00}{}^2} + \frac{1}{2}\frac{\phi_{02}}{\phi_{00}} - \right. \\ & - \left.\left.\frac{1}{2}\frac{\eta H_{u00}}{\sqrt{\phi_{00}}}\right]v^2 - \frac{\eta H_{w00}}{\sqrt{\phi_{00}}}uv' + \frac{1}{2}(u'^2 + v'^2)\right\} = \\ & = \sqrt{\phi_{00}}\,\{B_1 u^2 + B_2 uv + B_3 v^2 + B_4 uv' + \tfrac{1}{2}(u'^2 + v'^2)\}, \\ F_3 = {} & \sqrt{\phi_{00}}\left\{\left[\frac{1}{2}\frac{\phi_{30}}{\phi_{00}} - \frac{1}{4}\frac{\phi_{20}\phi_{10}}{\phi_{00}{}^2} + \frac{1}{2}\frac{\phi_{20}}{\rho\phi_{00}} + \frac{1}{16}\frac{\phi_{10}{}^3}{\phi_{00}{}^3} - \frac{1}{8}\frac{\phi_{10}{}^2}{\rho\phi_{00}{}^2} + \right.\right. \\ & + \left.\frac{1}{4}\frac{\phi_{10}\tau^2}{\phi_{00}} - \frac{1}{2}\frac{\tau^2}{\rho} + \frac{1}{6}\frac{\eta H_{v00}}{\sqrt{\phi_{00}}}\tau^2 - \frac{1}{3}\frac{\eta H_{v10}}{\rho\sqrt{\phi_{00}}} - \frac{1}{3}\frac{\eta H_{v20}}{\sqrt{\phi_{00}}}\right]u^3 + \\ & + \left[\frac{1}{2}\frac{\phi_{21}}{\phi_{00}} - \frac{1}{4}\frac{\phi_{20}\phi_{01}}{\phi_{00}{}^2} - \frac{1}{4}\frac{\phi_{11}\phi_{10}}{\phi_{00}{}^2} + \frac{1}{2}\frac{\phi_{11}}{\rho\phi_{00}} + \frac{3}{16}\frac{\phi_{10}{}^2\phi_{01}}{\phi_{00}{}^3} - \right. \\ & - \frac{1}{4}\frac{\phi_{10}\phi_{01}}{\rho\phi_{00}{}^2} + \frac{1}{4}\frac{\tau^2\phi_{01}}{\phi_{00}} - \frac{\eta H_{u20}}{\sqrt{\phi_{00}}} - \frac{1}{2}\frac{H_{w10}{}'}{\sqrt{\phi_{00}}} - \frac{\eta H_{u10}}{\rho\sqrt{\phi_{00}}} - \\ & - \left.\frac{1}{2}\frac{\eta H_{u00}}{\sqrt{\phi_{00}}}\tau^2\right]u^2 v + \left[\frac{1}{2}\frac{\phi_{12}}{\phi_{00}} - \frac{1}{4}\frac{\phi_{11}\phi_{01}}{\phi_{00}{}^2} - \frac{1}{4}\frac{\phi_{02}\phi_{10}}{\phi_{00}{}^2} + \right. \end{aligned}\right\} \tag{69}$$

$$\begin{aligned}
&+\frac{1}{2}\frac{\phi_{02}}{\rho\phi_{00}}+\frac{3}{16}\frac{\phi_{10}\phi_{01}{}^2}{\phi_{00}{}^3}-\frac{1}{8}\frac{\phi_{01}{}^2}{\rho\phi_{00}{}^2}+\frac{1}{4}\frac{\phi_{10}\tau^2}{\phi_{00}}-\frac{\tau^2}{2\rho}-\\
&-\frac{1}{2}\frac{\eta H_{u11}}{\sqrt{\phi_{00}}}-\frac{1}{2}\frac{\eta H_{w01}}{\sqrt{\phi_{00}}}-\frac{1}{2}\frac{\eta H_{u01}}{\rho\sqrt{\phi_{00}}}\Bigg]uv^2+\\
&+\Bigg[\frac{1}{2}\frac{\phi_{03}}{\phi_{00}}-\frac{1}{4}\frac{\phi_{02}\phi_{01}}{\phi_{00}{}^2}+\frac{1}{16}\frac{\phi_{01}{}^3}{\phi_{00}{}^3}+\frac{1}{4}\frac{\phi_{01}\tau^2}{\phi_{00}}-\frac{1}{3}\frac{\eta H_{u02}}{\sqrt{\phi_{00}}}-\\
&-\frac{1}{6}\frac{\eta H_{u00}}{\sqrt{\phi_{00}}}\tau^2\Bigg]v^3+\left[\frac{1}{4}\frac{\phi_{10}}{\phi_{00}}-\frac{1}{2\rho}\right]u(u'^2+v'^2)+\\
&+\frac{1}{4}\frac{\phi_{01}}{\phi_{00}}v(u'^2+v'^2)-\frac{1}{2}\frac{\eta H_{w10}}{\sqrt{\phi_{00}}}u^2v'-\frac{\eta H_{w01}}{\sqrt{\phi_{00}}}uvv'\Bigg\}=\\
&=\sqrt{\phi_{00}}\,\{C_1u^3+C_2u^2v+C_3uv^2+C_4v^3+\\
&\qquad+C_5u(u'^2+v'^2)+C_6v(u'^2+v'^2)+C_7u^2v'+C_8uvv'\}.
\end{aligned}\tag{69}$$

The symbol φ was chosen for the electrostatic potential and the coefficients in the series (63) were called $\phi_{m,n}$. The magnetic scalar potential was called ψ_M. Using the relations $\vec{H} = -\nabla\psi_M$ and $\vec{H} = \nabla \times \vec{A}$ it is possible to derive series expressions for the components of $\vec{H}$ and $\vec{A}$. The coefficients of such series for H_u, H_v and H_w are called H_{umn}, H_{vmn}, H_{wmn}. The corresponding Euler-Lagrange equations belonging to the Hamilton principle are

$$\frac{d}{dw}\left(\frac{\partial F}{\partial u'}\right)=\frac{\partial F}{\partial u},\qquad \frac{d}{dw}\left(\frac{\partial F}{\partial v'}\right)=\frac{\partial F}{\partial v}.\tag{70}$$

As usual, the term F_0 does not contribute to the motion of electrons; the term F_1 leads to an expression for the curvature of the optical axis. In case the optical axis was initially chosen as a possible electron path (verified by direct integration of the equation of motion), the equations of motion resulting from F_1 will be identities. The equations are

$$\left.\begin{aligned}
&\frac{1}{\rho}+\frac{\phi_{10}}{2\phi_{00}}+\frac{\eta H_{v00}}{\sqrt{\phi_{00}}}=0,\\
\text{and}\qquad&\\
&\frac{\phi_{01}}{2\phi_{00}}-\frac{\eta H_{u00}}{\sqrt{\phi_{00}}}=0.
\end{aligned}\right\}\tag{71}$$

Ordinarily, the coefficients ϕ_{01} and H_{u00} will be equal to zero.

If the term F_2 is substituted in the Euler-Lagrange equations, the following equations are obtained:

$$\left.\begin{aligned}\frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,u'\} &= \left(\tau^2+\frac{\phi_{20}}{\phi_{00}}+\frac{\phi_{10}}{\rho\phi_{00}}-\frac{1}{4}\frac{\phi_{10}{}^2}{\phi_{00}{}^2}+\frac{\eta H_{v10}}{\sqrt{\phi_{00}}}+\right.\\ &\quad\left.+\frac{\eta H_{v00}}{\rho\sqrt{\phi_{00}}}\right)u+\left(\frac{\phi_{11}}{2\phi_{00}}-\frac{\phi_{01}\phi_{10}}{4\phi_{00}{}^2}+\right.\\ &\quad\left.+\frac{\phi_{01}}{2\rho\phi_{00}}+\frac{\eta H_{v01}}{\sqrt{\phi_{00}}}\right)v-\frac{\eta H_{w00}}{\sqrt{\phi_{00}}}v',\\ \frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,v'\} &= \left(\frac{\phi_{11}}{2\phi_{00}}-\frac{\phi_{01}\phi_{10}}{4\phi_{00}{}^2}-\frac{\eta H_{u10}}{\sqrt{\phi_{00}}}\right)u+\\ &\quad+\left(\frac{\phi_{02}}{\phi_{00}}-\frac{\phi_{01}{}^2}{4\phi_{00}{}^2}+\tau^2-\frac{\eta H_{u01}}{\sqrt{\phi_{00}}}\right)v+\frac{\eta H_{w00}}{\sqrt{\phi_{00}}}u'.\end{aligned}\right\}\qquad(72)$$

The solutions of these equations describe the so-called first order electron path, in electron optics usually called the gaussian path. Substituting the term F_3 in the Euler-Lagrange equations, one obtains

$$\left.\begin{aligned}&\frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,u'\}-(2B_1u+B_2v+B_4v')=\{3C_1u_g{}^2+\\ &\quad+2C_2u_gv_g+C_3v_g{}^2+C_5(u_g'^2+v_g'^2)+2C_7u_gv_g'+C_8v_gv_g'\}-\\ &\qquad-\frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,[2C_5u_gu_g'+2C_6v_gu_g']\},\\ &\frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,v'\}-(B_2u+2B_3v-(\sqrt{\phi_{00}}\,B_4u)')=\\ &=\{C_2u_g{}^2+2C_3u_gv_g+3C_4v_g{}^2+C_6(u_g'^2+v_g'^2)+C_8u_gv_g'\}-\\ &\quad-\frac{1}{\sqrt{\phi_{00}}}\frac{d}{dw}\{\sqrt{\phi_{00}}\,[2C_5u_gv_g'+2C_6v_gv_g'+C_7u_g{}^2+C_8u_gv_g]\},\end{aligned}\right\}\qquad(73)$$

where the coefficients $B_{m,n}$, $C_{m,n}$ are defined by eq. (69). Here the solution of the first order electron path was substituted in all terms higher than the first degree. The solution of this system of differential equations obtained describes the electron path to the next higher degree of accuracy.

2. A Crossed Field Mass Spectrometer with a Radial Electrical Field

As the first application of the general theory developed so far, a special cross-field mass spectrometer[17,18,19] will be discussed, the π-radian mass spectrometer will also be mentioned.

The deflection field is assumed to be an electric field produced by concentric cylinders and a uniform magnetic field in the direction of the axis of these cylinders. The electrostatic potential is given by

$$\varphi(r)=\phi_0+\frac{V_2-V_1}{\ln R_2/R_1}\ln\frac{r}{\rho}.\qquad(74)$$

The magnetic field has only one component, namely:

$$H_z = H = \text{const.} \tag{75}$$

The curved optical axis is taken as the circle of radius ρ. By inspection of the equations of motion, it can be seen that such an electron path is possible if

$$\rho = \frac{(V_2 - V_1)/\ln R_2/R_1 + 2\phi_0}{\left(\frac{2e}{m}\phi_0\right)^{\frac{1}{2}}}\,\frac{1}{H}. \tag{76}$$

The potential series in the u, v, w system of coordinates is given by

$$\varphi(u, v, w) = \phi_0 + \frac{V_2 - V_1}{\ln R_2/R_1}\left(\frac{u}{\rho} - \frac{1}{2}\frac{u^2}{\rho^2} + \frac{1}{3}\frac{u^3}{\rho^3} - \frac{1}{4}\frac{u^4}{\rho^4} + - \cdots\right). \tag{77}$$

With the abbreviation:

$$y = -\frac{1}{2\phi_0}(V_2 - V_1)/\ln R_2/R_1, \tag{78}$$

one obtains

$$\left.\begin{aligned} \phi_{00} &= \phi_0, \\ \phi_{10} &= -\frac{2\phi_0}{\rho}\cdot y, \\ \phi_{20} &= \frac{\phi_0}{\rho^2}y, \\ \phi_{30} &= -\frac{2}{3}\frac{\phi_0}{\rho^3}y. \end{aligned}\right\} \tag{79}$$

The magnetic field has only one component H_{v00}:

$$H_{v00} = -H. \tag{80}$$

Furthermore, $w = \rho\theta$ hence $\frac{d}{d\theta} = \rho\frac{d}{dw}$.

The zero-order path equations are identically satisfied if eq. (76) is taken into account:

$$\frac{1}{\rho} = -\frac{\phi_{10}}{2\phi_{00}} + \frac{\sqrt{\frac{e}{2m}}H_{v00}}{\sqrt{\phi_{00}}} = \frac{1}{\rho}y + \frac{1}{\sqrt{\phi_0}}\,\frac{\sqrt{\frac{e}{2m}}\,\frac{1}{\rho}[-2\phi_0 y + 2\phi_0]}{\sqrt{\frac{e}{2m}}\sqrt{\phi_0}} = \frac{1}{\rho}.$$

The first-order path equations become

$$\left.\begin{aligned} \frac{d}{d\theta}(u') &= -(1 + y^2)u, \\ \text{and} \qquad \frac{d}{d\theta}(v') &= 0, \end{aligned}\right\} \tag{81}$$

the solutions of which are

$$\left.\begin{aligned} u &= a \cos \sqrt{1 + y^2}\,\theta + b \sin \sqrt{1 + y^2}\,\theta, \\ \text{and} \qquad v &= c + d\theta. \end{aligned}\right\} \tag{82}$$

If $u = 0$ at $\theta = 0$, u will be equal to zero again at

$$\theta = \frac{1}{\sqrt{1 + y^2}}\pi.$$

For a purely magnetic field ($y = 0$) one has $\theta = \pi = 180°$, while for a purely electric field between concentric cylinders ($y = 1$),

$$\theta = \frac{\pi}{\sqrt{2}} = 127°17'.$$

The second-order aberrations are described by means of the term F_3 of eq. (69). The differential equation for u is:

$$\frac{d}{d\theta}(u') + (1 + y^2)u = (1 - y)\frac{d}{d\theta}(uu') + \frac{1}{2}y(1 - 3y^2)u^2 - \frac{1}{2}(1 + y)(u'^2 + v'^2), \tag{83}$$

which is correct for ions which have no initial slope in the v direction. The integration of this equation leads to an expression for the line width L in the image plane, which is

$$L = -\alpha^2 \frac{1 + \frac{1}{3}y + y^2 + 3y^3}{(1 + y^2)^2} \tag{84}$$

where α is the angular aperture in the object plane. In the case of a purely magnetic field one has

$$L_M = -\alpha^2 r_0,$$

and for an electrostatic field

$$L_e = -\tfrac{4}{3}\alpha^2 r_0.$$

This line width is further increased due to electrons which have an initial slope in the v direction.

It was shown that the focusing action of the deflection fields discussed is imperfect in the uw-plane and nonexistent in the vw-plane. Only a fraction of the current leaving the source is collected by an exit slit. Attempts have been made to improve mass spectrometer fields in two ways. Fields have been designed for providing perfect focusing in the uw-plane disregarding, however, motion of ions outside this plane. Examples of such fields were described by Bleakney and Hipple,[20] and Coggeshall.[21] Spectrometers have also been designed providing focusing action in both the u and v direction. Such instruments are called two-directional focusing spectrometers, and were described by Siegbahn and Svartholm,[22,24] Svartholm,[23] and Shull and Dennison.[25]*

3. A Crossed Field Mass Spectrometer with a Constant Electric Field

As an example of the perfect-focusing type of instrument, the mass spectrometer of Bleakney and Hipple[20] will be discussed. The common cycloid is chosen as the optical axis. The equations of this curve are

$$x = a(\theta - \sin\theta), \qquad y = a(1 - \cos\theta). \tag{85}$$

The line element of this curve is

$$dw = 2a \sin \tfrac{1}{2}\theta d\theta, \tag{86}$$

and the length of the arc of the common cycloid is, therefore,

$$w = \int_0^\theta 2a \sin \tfrac{1}{2}\theta d\theta = 4a[1 - \cos\theta]. \tag{87}$$

The slope is given by

$$\frac{dy}{dx} = \cos\frac{1}{2}\theta. \tag{88}$$

The equations of the normal are

$$\left.\begin{aligned} x &= a(\theta_0 - \sin\theta_0) + \cos\tfrac{1}{2}\theta_0 \cdot u, \\ y &= a(1 - \cos\theta_0) - \sin\tfrac{1}{2}\theta_0 \cdot u. \end{aligned}\right\} \tag{89}$$

The potential distribution is assumed to be

$$\varphi = \frac{V_2 - V_1}{2d} y. \tag{90}$$

---

* Shull and Dennison have called these instruments "double-focusing" spectrometers. It should be pointed out that such a term is misleading since it implies that the instrument focuses particles having different e/m ratios, velocities, and direction with respect to either two of the three different properties.

The expansion of this potential function in powers of u is

$$\varphi = \frac{V_2 - V_1}{2d}\left[2a \sin^2 \frac{1}{2}\theta - \sin \frac{1}{2}\theta \cdot u\right]. \tag{91}$$

The coefficients $\phi_{m,n}$ are, therefore,

$$\left.\begin{aligned} \phi_{00} &= \frac{V_2 - V_1}{2d} 2a \left[1 - \left(1 - \frac{w}{4a}\right)^2\right], \\ \phi_{10} &= -\frac{V_2 - V_1}{2d} \sqrt{1 - \left(1 - \frac{w}{2a}\right)^2}, \end{aligned}\right\} \tag{92}$$

all other $\phi_{m,n}$ being equal to zero.

The magnetic field has only one component:

$$H_u \equiv 0, \qquad H_v = -H = \text{const}, \qquad H_w \equiv 0. \tag{93}$$

In order for the cycloid to be an electron path, a must be given by

$$a = \frac{m}{e} \frac{V_2 - V_1}{2d} \frac{1}{H^2}. \tag{94}$$

As before, the zero-order path equation turns out to be an identity.

$$\frac{1}{\rho} = -\frac{\phi_{10}}{2\phi_{00}} - \frac{\sqrt{\frac{e}{2m}}(-H)}{\sqrt{\phi_{00}}} = -\frac{1}{\rho} + \frac{2}{\rho} = \frac{1}{\rho}. \tag{95}$$

The first-order differential equations given by

$$\left.\begin{aligned} \frac{1}{\sqrt{\phi_{00}}} \frac{d}{dw}\left\{\sqrt{\phi_{00}}\, u'\right\} &= \left(\frac{\phi_{10}}{\rho\phi_{00}} - \frac{\phi_{10}{}^2}{4\phi_{00}{}^2} + \frac{\sqrt{\frac{e}{2m}} H_{v00}}{\rho\sqrt{\phi_{00}}}\right), \\ \frac{1}{\sqrt{\phi_{00}}} \frac{d}{dw}\left\{\sqrt{\phi_{00}}\, v'\right\} &= 0, \end{aligned}\right\} \tag{96}$$

may easily be transformed to

$$\left.\begin{aligned} \frac{d^2u}{d\theta^2} + \frac{1}{4}u &= 0, \\ \frac{d^2v}{d\theta^2} &= 0. \end{aligned}\right\} \tag{97}$$

The solutions of these equations are

$$u = A \cos \tfrac{1}{2}\theta + B \sin \tfrac{1}{2}\theta, \qquad v = A_1 + B_1\theta, \tag{98}$$

or with obvious initial conditions

$$u = u_0' \sin \tfrac{1}{2}\theta, \qquad v = v_0'\theta. \tag{99}$$

If u is equal to zero at $\theta = 0$ then u will be zero again at $\theta = 2\pi$. It has been shown that the focusing takes place at $\theta = 2\pi$.

In order to compute the aberrations, F_3 is evaluated, the result being

$$\left.\begin{aligned} F_3 &= \sqrt{\phi_{00}} \left\{ \left[\frac{1}{16} \frac{\phi_{10}^3}{\phi_{00}^3} - \frac{1}{8} \frac{\phi_{10}^2}{\rho\phi_{00}^2} \right] u^3 + \left[\frac{1}{4} \frac{\phi_{10}}{\phi_{00}} - \frac{1}{2\rho} \right] uu'^2 \right\} \\ &= \sqrt{\phi_{00}} \{ C_1 u^3 + C_5 uu'^2 \}. \end{aligned}\right\} \tag{100}$$

The perfect-focusing properties of the mass spectrometer field under discussion are expressed by stating that the field is free of aberrations (for electron paths without initial slope in the $\imath$ direction).

It can easily be shown that

$$C_1 = C_5 = 0. \tag{101}$$

Higher-order aberrations will not be computed.

4. The Determination of Curved Optical Axes

The results of investigations by Coggeshall[21] and Coggeshall and Muskat[26] may be used to determine a curved optical axis for a number of mass-spectrometer fields. These authors integrated the equations of motion of charged particles for single magnetic fields and combined magnetic and electric fields. The fields are not necessarily uniform but are restricted in that the magnetic and electric field-strength vectors are mutually perpendicular and are functions of one coordinate only; the electric vector has only one component in the direction of that axis on whose coordinate it depends as a function.

Let $\varphi(x)$ be the electric potential and $H_z(x)$ the magnetic field strength component. The equation of motion in the yx plane is then

$$\frac{d}{dx}\left[\sqrt{\frac{2e}{m}\varphi(x)}\,\frac{y'}{\sqrt{1+y'^2}}\right] = \frac{e}{m} H_z(x), \tag{102}$$

or

$$\sqrt{\frac{2e}{m}\varphi(x)}\,\frac{y'}{\sqrt{1+y'^2}} = \frac{e}{m}\left[\int_{x_0} H_z(x)dx + \bar{c}\right]. \tag{103}$$

Defining

$$\int_{x_0}^{x} H_z(x)dx + \bar{c} = f(x), \tag{104}$$

one obtains

$$\frac{dy}{dx} = y' = \pm \frac{f(x)}{\left[\dfrac{2m}{e}\varphi(x) - f^2(x)\right]^{\frac{1}{2}}}. \tag{105}$$

The corresponding equation for the case of a radial electric field given by an electric potential function $\varphi(r)$ and a magnetic field described by a function $H_z(r)$ is

$$\frac{d\theta}{dr} = \pm \frac{1}{r} \frac{g(r)}{\left[\frac{2m}{e}\varphi(r) - g^2(r)\right]^{\frac{1}{2}}}, \tag{106}$$

where $g(r)$ is defined by

$$g(r) = \frac{1}{r} \int_{r_0}^{r} \frac{1}{r} H_z(r) dr + \frac{\bar{c}}{r}. \tag{107}$$

Due to the assumptions concerning the orthagonality and functional dependency of the electric and magnetic field vectors, the above integrations apply only to motion in a median plane.* In other regions of actual fields these assumptions are generally not satisfied.

For a number of simple fields, eqs. (105) and (106) can be integrated in terms of elementary functions. Numerical methods may be employed if analytical methods fail or if the fields are given experimentally.

Fields investigated in this manner are:

1. Homogeneous magnetic field.[26]
2. Linearly varying magnetic field.[26]
3. Exponentially varying magnetic field.[26]
4. Radial fields.[26]
5. Homogeneous electric and magnetic fields.[20,21]
6. Linearly varying electric field and homogeneous magnetic field.[21]
7. Quadratically varying electric field and homogeneous magnetic fields.[21]
8. Exponentially varying electric and magnetic fields.[21]
9. Homogeneous magnetic field and an electric field varying inversely with the radius.[21]
10. Homogeneous magnetic field and an electric field varying directly with the radius.[21]

All fields with the exception of No. 9 admit of analytic solutions.

Some of these fields provide perfect focusing in the median plane, which is a desirable feature. The investigations may be supplemented by means of the theory outlined above in order to obtain information about ion or electron paths other than those in the median plane. This information is necessary to arrive at a correct basis for the comparison of the relative merits of the perfect-focusing mass spectrometer and the two-directional focusing mass spectrometer. The theory of the latter type instrument is now being developed.

---

* Such a plane is located symmetrically relative to the magnetic pole pieces and located properly relative to electrodes producing the electric field.

5. Two-Directional Focusing with Deflection Type Fields

The first-order electron path was given by eq. (72). All electrons leaving a point in the object plane $w = w_0$ go through the point in the image plane $w = w_i$, if there are particular solutions in u and v which are equal to zero at $w = w_i$. A study of the path differential equations with the object of obtaining field distributions for which eq. (72) has this property is rather complicated. The investigations by Wendt[16] were, therefore, restricted to the case of rotational symmetry about the curved optical axis. In this case, the solutions u and v are of the same functional type. The first order path differential equations must be of the form

$$\left.\begin{aligned} \frac{d}{dw}(nu' - mv) &= pu + mv', \\ \frac{d}{dw}(nv' + mu) &= pv - mu', \end{aligned}\right\} \tag{108}$$

or

$$\left.\begin{aligned} \frac{d}{dw}(nu') &= pu + m'v + 2mv', \\ \frac{d}{dw}(nv') &= pv - m'u - 2mu'. \end{aligned}\right\} \tag{109}$$

Comparing corresponding terms in eqs. (72) and (109) one obtains:

$$\left.\begin{aligned} \frac{p}{\sqrt{\phi_{00}}} &= \tau^2 + \frac{\phi_{20}}{\phi_{00}} + \frac{\phi_{10}}{\rho\phi_{00}} - \frac{\phi_{10}{}^2}{4\phi_{00}{}^2} + \frac{\eta H_{v10}}{\sqrt{\phi_{00}}} + \frac{\eta H_{v00}}{\rho\sqrt{\phi_{00}}} \\ &= \tau^2 + \frac{\phi_{02}}{\phi_{00}} - \frac{\phi_{01}{}^2}{4\phi_{00}{}^2} - \frac{\eta H_{u01}}{\sqrt{\phi_{00}}}, \\ \frac{m'}{\sqrt{\phi_{00}}} &= -\frac{\phi_{11}}{2\phi_{00}} - \frac{\phi_{01}\phi_{10}}{4\phi_{00}{}^2} + \frac{\phi_{01}}{2\rho\phi_{00}} + \frac{\eta H_{v01}}{\sqrt{\phi_{00}}} \\ &= -\frac{\phi_{11}}{2\phi_{00}} + \frac{\phi_{01}\phi_{10}}{4\phi_{00}{}^2} + \frac{\eta H_{u10}}{\sqrt{\phi_{00}}} = -\frac{\eta}{2}\frac{H_{w00}{}'}{\sqrt{\phi_{00}}}. \end{aligned}\right\} \tag{110}$$

Relations between the coefficients of the potential series are obtained by substituting these series into LaPlace's equation. These relations, combined with eq. (110) give:

$$\left.\begin{aligned} \frac{p}{\sqrt{\phi_{00}}} &= \tau^2 - \frac{\phi_{10}{}^2}{4\phi_{00}{}^2} - \frac{\phi_{02}{}^2}{8\phi_{00}{}^2} - \frac{\phi_{00}{}''}{4\phi_{00}} + \frac{\eta H_{v00}\phi_{10}}{2\phi_0{}^{3/2}} - \frac{\eta^2 H_{v00}{}^2}{2\phi_{00}}, \\ \frac{m}{\sqrt{\phi_{00}}} &= -\frac{\eta}{2}\frac{H_{w00}}{\sqrt{\phi_{00}}}. \end{aligned}\right\} \tag{111}$$

The equations in u and v may be combined into one single equation by writing

$$\bar{\sigma} = \bar{u} + i\bar{v} = (u + iv)e^{2\chi(w)}, \qquad \bar{w} = w. \tag{112}$$

The magnetic component H_{w00} causes a rotation of the image. The angle χ is given by

$$\chi = -\int_{w_0}^{w} \frac{m}{\sqrt{\phi_{00}}}\, dw. \tag{113}$$

The differential equation in $\bar{\sigma}$ is given by

$$\frac{d}{dw}\left(\sqrt{\phi_{00}}\,\bar{\sigma}'\right) = \left(p - \frac{m^2}{\sqrt{\phi_{00}}}\right)\bar{\sigma}. \tag{114}$$

The solution of this equation may be written in the form

$$\bar{\sigma} = \bar{a}g(w) + \bar{b}h(w). \tag{115}$$

It may be verified that the solutions for u and v can be written in the form

$$\left.\begin{aligned} u &= [(u_0 \cos\chi - v_0 \sin\chi)g + (u_a \cos(\chi - \chi_i) - \\ &\qquad\qquad - v_a \sin(\chi - \chi_i))h], \\ v &= [(u_0 \sin\chi + v_0 \cos\chi)g + (u_a \sin(\chi - \chi_i) + \\ &\qquad\qquad + v_a \cos(\chi - \chi_i))h]. \end{aligned}\right\} \tag{116}$$

The particular solutions g and h must satisfy the following boundary conditions:

$$\left.\begin{aligned} h(w_0) &= 0, & h(w_a) &= 1, \\ g(w_0) &= 1, & g(w_a) &= 0. \end{aligned}\right\} \tag{117}$$

$w = w_a$ is the coordinate of an aperture plane. The image plane is located at $w = w_i$, where

$$h(w_i) = 0. \tag{118}$$

It is assumed that the space between the planes $w = w_a$ and $w = w_i$ is field-free.

The dioptrics of focusing systems of arbitrarily curved axes may be patterned after that of ordinary focusing systems. Relations between object and image distances, magnification, focal lengths, and location of the principal planes will be the same in both types of systems.

It has been seen that it is necessary to know the particular solution of eq. (114) in order to be able to determine focal lengths and the locations of the principal planes. These solutions must satisfy the conditions

$$\bar{\sigma}(w_0) = 1, \qquad \bar{\sigma}'(w_0) = 0. \tag{119}$$

The focal length on the image side is then given by

$$1/f_i = -\boldsymbol{\sigma}_i'. \tag{120}$$

The focal length on the object side is

$$1/f_0 = -1/f_i \sqrt{V_i/V_0}, \tag{121}$$

where V_0 and V_i are the potentials in the object and image planes respectively. The location of the principal plane of the image side is given by

$$w_{Hi} = w_i - f_i(1 - \boldsymbol{\sigma}_i). \tag{122}$$

The location of the principal plane of the object side is

$$w_{H_0} = w_0 - f_0(1 - \boldsymbol{\sigma}_0). \tag{123}$$

The theory of lenses of rotational symmetry uses methods of successive approximation to obtain a particular solution of the path differential equations. As the first solution, $\boldsymbol{\sigma}_1 = 1 = \text{const}$ is ordinarily chosen. The second integration gives

$$\boldsymbol{\sigma}_2 = 1 + \int_{w_0}^{w} \frac{dw}{\sqrt{\phi_{00}}} \int_{w_0}^{w} \left(p - \frac{m^2}{\sqrt{\phi_{00}}}\right) dw. \tag{124}$$

Repeated substitution of solutions into the right side of eq. (114) yields improved solutions.

In the case of arbitrarily curved optical axes, however, it is often true that the electron path has the value zero within the field, hence, $\boldsymbol{\sigma}_1 = \text{const}$ taken as the first approximation, appears undesirable. If it is assumed that ϕ_{00} and $\left(p - \frac{m^2}{\sqrt{\phi_{00}}}\right)$ change only gradually so that one considers them as practically constant within an interval, eq. (114) has as a solution a circular function. As a first step, therefore, one may assume

$$\boldsymbol{\sigma}_1 = \cos\left(\int_{w_0}^{w} \alpha dw\right), \tag{125}$$

where

$$\alpha = \sqrt{\frac{m^2}{\phi_{00}} - \frac{p}{\sqrt{\phi_{00}}}}. \tag{126}$$

The following approximation may then be written:

$$\boldsymbol{\sigma}_2 = 1 + \int_{w_0}^{w} \frac{dw}{\sqrt{\phi_{00}}} \int_{w_0}^{w} \sqrt{\phi_{00}}\, \alpha^2 \cos\left(\int_{w_0}^{w} \alpha dw\right) dw. \tag{127}$$

To this degree of approximation (assuming $w = w_0$ and $w = w_i$ in field-free space), one may obtain for the focal lengths and the location of the principal planes

$$\frac{1}{f_i} = -\frac{1}{f_0}\sqrt{\frac{V_0}{V_i}} = -\frac{1}{\sqrt{V_i}}\int_{-\infty}^{+\infty} \sqrt{\phi_{00}}\,\alpha^2 \cos\left(\int_{-\infty}^{+\infty} \alpha dw\right) dw, \tag{128}$$

$$\left.\begin{aligned} w_{H_0} &= f_0\left[\int_{-\infty}^{+\infty} w\alpha^2 \cos\left(\int_{w}^{\infty} \alpha dw\right) dw + \right. \\ &\qquad \left. + \int_{-\infty}^{+\infty} \frac{w\phi_{00}'dw}{\phi_{00}^{\frac{3}{2}}} \int_{w}^{\infty} \alpha^2 \sqrt{\phi_{00}} \cos\left(\int_{w}^{\infty} \alpha dw\right) dw\right], \\ w_{H_i} &= -f_i\left[\int_{-\infty}^{+\infty} w\alpha^2 \cos\left(\int_{-\infty}^{w} \alpha dw\right) dw + \right. \\ &\qquad \left. + \int_{-\infty}^{+\infty} \frac{w\phi_{00}'dw}{\phi_{00}^{\frac{3}{2}}} \int_{-\infty}^{w} \alpha^2 \sqrt{\phi_{00}} \cos\left(\int_{-\infty}^{w} \alpha dw\right) dw\right]. \end{aligned}\right\} \tag{129}$$

6. *Purely Electric Deflection Fields*

In this case the following set of relations is obtained:

$$\phi_{01} = \phi_{11} = 0, \qquad \frac{1}{\rho} = -\frac{\phi_{10}}{2\phi_{00}}, \qquad \frac{\phi_{20}}{\phi_{00}} = \frac{\phi_{10}^2}{2\phi_{00}^2} - \frac{\phi_{00}''}{4\phi_{00}},$$
$$\frac{\phi_{02}}{\phi_{00}} = -\frac{\phi_{10}^2}{4\phi_{00}^2} - \frac{\phi_{00}''}{4\phi_{00}}. \tag{130}$$

The potential series has, therefore, the form

$$\varphi = \phi_{00} + \phi_{10}u + \left(\frac{\phi_{10}^2}{2\phi_{00}} - \frac{\phi_{00}''}{4}\right)w^2 - \left(\frac{\phi_{10}^2}{4\phi_{00}} + \frac{\phi_{00}''}{4}\right)v^2 + \cdots, \tag{131}$$

or

$$\frac{\varphi}{\phi_{00}} = 1 - \frac{2u}{\rho} + \left(\frac{2}{\rho^2} - \frac{\phi_{00}''}{4\phi_{00}}\right)u^2 - \left(\frac{1}{\rho^2} + \frac{\phi_{00}''}{4\phi_{00}}\right)v^2 + \cdots. \tag{132}$$

From these considerations, the equipotential line distribution is given by

$$\frac{\left[\dfrac{u}{\rho} - \dfrac{1}{2\left(1 - \dfrac{\phi_{00}''\rho^2}{8\phi_{00}}\right)}\right]^2}{\dfrac{1 + 2\left(\dfrac{\varphi_k}{\phi_{00}} - 1\right)\left(1 - \dfrac{\phi_{00}''\rho^2}{8\phi_{00}}\right)}{4\left(1 - \dfrac{\phi_{00}''\rho^2}{8\phi_{00}}\right)^2}} - \frac{\left(\dfrac{u}{\rho}\right)^2}{\dfrac{1 + 2\left(\dfrac{\varphi_k}{\phi_{00}} - 1\right)\left(1 - \dfrac{\phi_{00}''\rho^2}{8\phi_{00}}\right)}{2\left(1 - \dfrac{\phi_{00}''\rho^2}{8\phi_{00}}\right)\left(1 + \dfrac{\phi_{00}''\rho^2}{4\phi_{00}}\right)}} = 1, \tag{133}$$

which is the equation for a set of confocal hyperbolas. If the potential on the w-axis is constant, that is

$$\phi_{00}'' = 0,$$

we obtain

$$\frac{\left(\frac{u}{\rho} - \frac{1}{2}\right)}{\frac{1}{2}\left(\frac{\varphi_k}{\phi_{00}} - \frac{1}{2}\right)} - \frac{\left(\frac{u}{\rho}\right)^2}{\frac{\varphi_k}{\phi_{00}} - \frac{1}{2}} = 1. \tag{134}$$

This potential distribution is shown in Fig. 13 which also shows electrodes producing such a field.

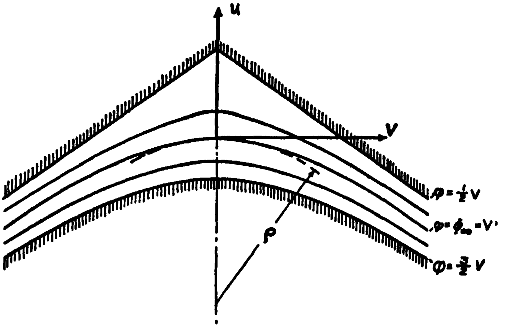


FIG. 13.—Electrodes of an electrical focusing system with a curved axis. (From *Z. Phys.*, **120,** No. 11/12, 1943, printed by permission of the Alien Property Custodian.)

One of the electrodes was assumed to coincide with the asymptotes of the set of confocal hyperbolas. The equipotential line passing through the point $u = 0$, $v = 0$ was assumed to be at the potential $\varphi_k = \phi_{00} = V$. The second electrode was taken to coincide with the equipotential line $\varphi = \frac{3}{2}V$.

7. *Purely Magnetic Deflection Field*

In this case,

$$\left.\begin{aligned} \phi_{00} &= \text{const}, \\ H_{u00} &= 0, \\ H_{u10} &= H_{v01}' = -\tfrac{1}{2}H_{w00}', \\ H_{u01} &= H_{V10} = \frac{\eta}{2}\frac{H_{v00}^2}{\sqrt{\phi_{00}}}, \\ \frac{1}{\rho} &= -\frac{\eta H_{v00}}{\sqrt{\phi_{00}}}. \end{aligned}\right\} \tag{135}$$

Introducing the magnetic scalar potential ψ_M, the following series is obtained for this quantity

$$p\psi_M = -\int pH_{w00}dw + \frac{v}{\rho} + \frac{p}{4}H_{w00}'(u^2 + v^2) - \frac{uv}{2\rho^2}, \tag{136}$$

where

$$p = \frac{\eta}{\sqrt{\phi_{00}}}. \tag{137}$$

This equation represents a set of ellipses, parabolas, or hyperbolas depending on whether, in a plane $w = \text{const}$, $p^2H_{w00}'^2$ is greater, equal, or smaller than $\frac{1}{4}\rho^4$.

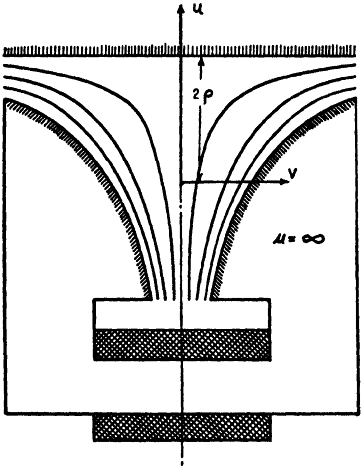


FIG. 14.—Pole pieces of a magnetic focusing system with a curved axis. Equipotential lines of the magnetic scalar potential are shown. (From *Z. Phys.*, **120,** No. 11/12, 1943, by permission of the Alien Property Custodian.)

In the case

$$H_{w00} = 0, \tag{138}$$

we obtain

$$p\psi_M = \frac{v}{\rho} - \frac{uv}{2\rho^2} + \cdots \approx -\frac{v}{\rho}\left(\frac{u}{2\rho} - 1\right). \tag{139}$$

This equation describes the set of equilateral hyperbolas with the asymptotes

$$u = 2\rho, \qquad v = 0. \tag{140}$$

Pole pieces producing such a magnetic field are shown in Fig. 14. The magnetic field distribution discussed by Shull and Dennison[25] is a special case of the one just described.

REFERENCES

1. Picht, J., and Himpan, J. *Ann. Phys. Lpz.*, **39**, 409 (1941); **43**, 53 (1943).
2. Wendt, G. *Die Telefunkenrohre*, **15**, 100 (1939); *Z. Phys.*, **118**, 593 (1942).
3. Hutter, R. G. E. *J. Appl. Phys.*, **18**, 740–758 (1947); **18**, 797–810 (1947).
4. Hutter, R. G. E. *Proc. Nat. Electronics Conf., Chicago*, 424–453 (1946).
5. Cazalas, A. *Ônde Élect.*, **26**, 467–471 (1946).
6. Rüdenberg, H. G. *J. Appl. Phys.*, **16**, 279 (1945).
7. Hutter, R. G. E. *J. Appl. Phys.*, **18**, 797–810 (1947).
8. Moss, H. *J. Televis. Soc.*, **4**, 206 (1946).
9. Schlesinger, K. U.S. Pat. No. 2,227,036 (1940).
10. Schlesinger, K. U.S. Pat. No. 2,227,020 (1940).
11. Brüche, E., and Henneberg, W. U.S. Pat. No. 2,101,669 (1937).
12. Fleming-Williams, B. D. *Wireless Engr.*, **17**, 61–64 (1940).
13. Sharpe, J. *Electronic Engng.*, **18**, 385 (1946).
14. Bowie, R. M. U.S. Pats. No. 2,211,613 and 2,211,614.
15. Oliphant, M. L., Shire, E. S., and Crowther, B. M. *Proc. Roy. Soc.*, **A1946**, 922 (1934).
16. Wendt, G. *Z. Phys.*, **120**, 720–740 (1943).
17. Hutter, R. G. E. *Phys. Rev.*, **67**, 248-253 (1945).
18. Herzog, R. *Z. Phys.*, **89**, 447 (1934).
19. Henneberg, W. *Ann. Phys. Lpz.*, **19**, 335 (1934); **20**, 1 (1934); **21**, 390 (1934).
20. Bleakney and Hipple. *Phys. Rev.*, **53**, 521 (1938).
21. Coggeshall, N. D. *Phys. Rev.*, **70**, 270–280 (1946).
22. Svartholm, N., and Siegbahn, K. *Ark. Mat. Astr. Fys.*, **33A**, No. 21 (1946).
23. Svartholm, N. *Ark. Mat., Astr. Fys.*, **33A**, No. 24 (1946).
24. Siegbahn, K., and Svartholm, N. *Nature Lond.*, **157**, 872 (1946).
25. Shull, F. B., and Dennison, D. M. *Phys. Rev.*, **71**, 681–687 (1947).
26. Coggeshall, N. D., and Muskat, M. *Phys. Rev.*, **66**, 187–198 (1944).

Modern Mass Spectroscopy

MARK G. INGHRAM

Argonne National Laboratory
Chicago 80, Illinois

CONTENTS

Page
I. Introduction.... 219
II. General Theory 220
  1. Ion Trajectories in Magnetic and Electrostatic Fields 221
  2. Focusing Properties of Magnetic and Electrostatic Fields 223
  3. Velocity Focusing Properties of a Combination of Electrostatic and Magnetic Fields 225
  4. Mass Dispersion Produced by a Magnetic Analyzer 226
  5. Line Widths Produced by Electrostatic and Magnetic Analyzers 226
III. Apparatus 227
  1. Ion Sources Used in Mass Spectroscopy 228
  2. Sample Handling Systems 232
  3. Mass Analyzers 237
  4. Direction Focusing Magnetic Analyzers 240
  5. Velocity Selection—Direction Focusing Analyzers 242
  6. Double Focusing Analyzers 243
  7. Velocity Selection Analyzers 247
  8. Ion Detection Systems 248
  9. Electronic Components 252
IV. Uses of the Mass Spectroscope 253
  1. Isotope Existence 253
  2. Isotopic Abundances 253
  3. Packing Fractions 256
  4. Determination of the Mass of Radioactive Isotopes 258
  5. Neutron Absorption Cross Sections 260
  6. Gas Analysis 260
  7. Solid Analysis 263
  8. Leak Detection 264
  9. Other Applications of the Mass Spectrometer 264
V. Commercially Available Mass Spectrometers 265
References 265

I. INTRODUCTION

Mass spectrographs and mass spectrometers are electronic instruments which analyze substances according to the mass of the constituent elements and molecules present.* The basic principle upon which they operate is that moving charged particles of the same energy or velocity, but differing in mass or charge are acted upon by forces of different magnitude while passing through magnetic fields. The particular arrangement of fields is a problem in ion optics. In general, the instruments consist of three major components; a source for producing a beam of ions; an analyzer into which the beam is projected and by means of which it is resolved into its characteristic mass components; and a detector system for recording the resolved ion beams.

The first important result due to the mass spectroscope was the discovery by J. J. Thomson[1] that neon was not a simple element, but that it consisted of a mixture of two different isotopes; later these were extended to three. Since that time 302 isotopes have been found that occur by natural processes in terrestrial matter.

The second important result due to the mass spectroscope was Aston's[2] proof that the masses of all isotopes are not simple multiples of a fundamental unit, but that they are characterized by a mass defect, that is, isotopes do not have integral masses. It is this mass defect that gives rise to the energy produced in fission.[3]

More recently the mass spectroscope has been applied in many other ways. These applications include the rapid analysis of hydrocarbon mixtures;[4–9] the use of tracers in biological, chemical, and metallurgical problems;[10] the most sensitive method known for locating gas leaks;[11] and the routine control of industrial plants.[12]

It is in the interest of acquainting the reader with this new and powerful tool that this article is written.

II. GENERAL THEORY

Unfortunately there is no such thing as a single mass spectrometer or spectrograph which can be called a universal machine, that is, one that can do all problems involved in mass spectroscopy. Generally, the mass spectrograph is employed for determining what masses are present or to determine mass defects, while the mass spectrometer is used to measure the relative abundances of the different components in the mass spectrum. There is, however, a great similarity in the components of the two types of machines, so that the apparatus will be discussed together.

---

* Mass spectrographs and mass spectrometers are two distinct types of instruments. The name mass spectrograph is generally restricted to those mass sensitive instruments which produce a focused mass spectrum on a photographic plate. The name mass spectrometer is applied to those machines which bring a focused beam of ions to a fixed collector, where it is measured electrically. The terms mass spectroscope and mass spectroscopy are used in a loose sense to include both types of machines, and will be so used in this article.

It will be sufficient for the purposes here to give a very brief summary of the important equations used in instrument design.

1. *Ion Trajectories in Magnetic and Electrostatic Fields*

When a charged particle moves through a magnetic field, a force directed perpendicularly to both the direction of motion and to the magnetic field is exerted upon it by that magnetic field. This force is represented by the vector equation:

$$\mathbf{F} = \frac{e\mathbf{v} \times \mathbf{B}}{c} \tag{1}$$

where **B** is the strength of the magnetic field in gauss, e is the charge on the moving particle in e.s.u., $\mathbf{v}$ is the velocity of the moving particle in cm./sec., and c is the velocity of light in cm./sec.

The ion path of a particle subjected to this force is obtained by integrating the differential equation formed by equating this force to the rate of change of momentum of the particle:

$$\mathbf{F} = \frac{e\mathbf{v} \times \mathbf{B}}{c} = \frac{d}{dt} m\mathbf{v} \tag{2}$$

where m is the mass of the particle in grams.

In the great majority of modern mass spectroscope designs the particles move at right angles to the magnetic field, and the velocity of these particles is very much less than the velocity of light. This equation can therefore be written in its nonvector nonrelativistic form:

$$F = ev\,(B/c) = mv^2/r_m \tag{3}$$

where r_m is the radius of curvature of the particle path in the magnetic field.

The kinetic energy of an ion moving with a velocity v is:

$$KE = \tfrac{1}{2}mv^2 = eV \tag{4}$$

where V is the voltage through which the ion was accelerated in e.s.u. Combining eqs. (3) and (4) to eliminate v:

$$r_m = (c/B)(2mV/e)^{\frac{1}{2}} \tag{5}$$

If the mass is expressed atomic mass units, the magnetic field in gauss, the radius of curvature in centimeters, and ion accelerating potentials in volts, eq. (5) becomes:

$$r_m = (144/B)(mV/e)^{\frac{1}{2}} \tag{6}$$

From this equation it can be seen, that the path an ion follows in the magnetic field is a function of the mass of that ion. It is this fact which is used as the basis of mass spectroscopy. Fig. 1 shows the trajectories in a magnetic field of two ion beams having the same energy, but differing in mass.

A similar derivation of ion deflection in an electrostatic field where the ion trajectories are at right angles to that field gives:

$$r_e = 2V/E \tag{7}$$

where r_e is the radius of curvature of the ion path in centimeters, E is the electrostatic field in e.s.u. volts/cm., and V is the voltage through which the ion was accelerated in e.s.u. From this equation it is apparent that the path an ion follows in an electrostatic field is not a function of the mass of that ion. The importance of electrostatic deflection in mass spectroscopy is that a combination of electrostatic and magnetic fields can be made, such that the velocity dispersion of the electrostatic field just compensates for the velocity dispersion of the magnetic field. Such a combination of fields gives a machine which operates independent of any inhomogeneity in the initial energy of the ion beam. This type of velocity correction is called "*velocity focusing.*" Several examples of mass spectrographs, which use velocity focusing, will be discussed in section III-6 of this article.

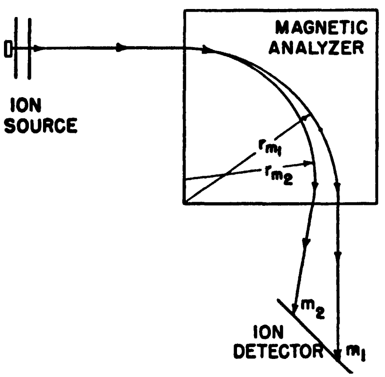


Fig. 1.—Diagram showing the trajectories in a magnetic field of two perfectly collimated ion beams having the same energy, but differing in mass.

It is to be noted in Fig. 1, that the deflection of the ion beams has been drawn as starting at the magnet face. Actually, the deflection starts before it reaches this face due to fringing effects. This fringing field maybe taken into account by assuming the effective field as larger than the pole faces by an amount equal to approximately one gap width.[13] This is, of course, only the first order correction since the path actually deviates before it reaches this point. However, in most applications it is pointless to make a detailed calculation of second order fringing corrections, since in all practical applications the final alignment must be determined experimentally.

Likewise, in the case of electrostatic fields a correction is necessary for the fringing effects due to the electrostatic field. These corrections have been worked out in detail by Rogers.[14]

2. Focusing Properties of Magnetic and Electrostatic Fields

Figure 1 shows the trajectories for perfectly collimated beams of ions. Theoretically, a perfectly collimated beam of ions is possible only with infinitely narrow slits. Such an ideal condition would reduce ion intensities to infinitely small values. In experimental applications a slit system in which the slits have finite width is necessary so that usable ion currents may be obtained. As a result, the ion beam is divergent as it enters the analyzing field. However, magnetic and electrostatic fields have focusing actions such that a divergent beam of ions of a single energy may be refocused. This type of focusing is called "*direction focusing.*"

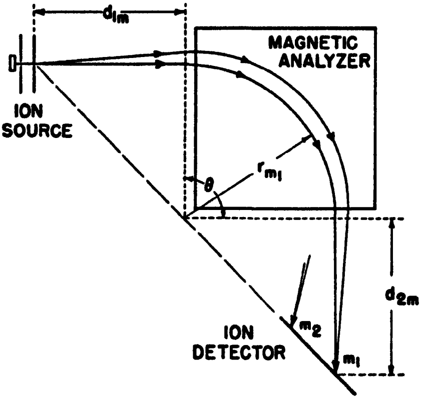


FIG. 2.—Diagram showing the direction focusing action of a magnetic field for ion beams homogeneous in energy.

The general problems of electrostatic and magnetic direction focusing have been solved by Herzog.[15] Referring to Fig. 2, the equation of direction focus for an ion beam of one mass, whose central beam enters and leaves the magnetic field at right angles to the faces of that field, can be written:

$$(d_{1m} - g_m)(d_{2m} - g_m) = f_m^2 \tag{8}$$

where f_m and g_m are defined by the equations:

$$f_m = r_m \csc \theta$$
$$g_m = r_m \cot \theta$$

d_{1m} is the distance from the apex of the divergent beam to the effective face of the magnetic field or object distance; d_{2m} is the distance from the effective exit face of the magnetic field to the point of focus or image distance; r_m is the radius of curvature of the particle path in the magnetic field; and θ is the angle through which the normal ion beam is deviated.

As an example of the application of eq. (8) consider the "symmetrical" type of mass analyzer with an ion deflection of 60°. The term "symmetrical" refers to the special case where d_{1m} is equal to d_{2m}. In this case we have from eq. (8):

$$d_{1m} = d_{2m} = r_m(\csc 60° + \cot 60°) \tag{9}$$

or:

$$d_{1m} = d_{2m} = 1.732r_m \qquad (10)$$

Thus, the object and image distances are 1.732 r_m units away from the effective faces of the magnet. The graphic construction of this particular focusing arrangement is shown in Fig. 5. This particular type of analyzer was used in conjunction with a direction focusing electrostatic analyzer by Bainbridge and Jordan[16] in 1936, it was used by itself as a mass spectrometer by Nier[17] in 1940.

A general statement which is valid for any angle of deflection in a magnetic field is that the defining slit of the source, the center of curvature of the central ion beam, and the final focal points are all in a straight line, see Figs. 2 and 5.

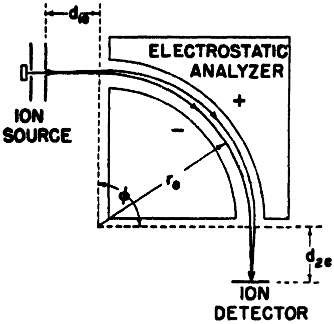


FIG. 3.—Diagram showing the direction focusing action of an electrostatic field for an ion beam homogeneous in energy.

Referring to Fig. 3, the equation for direction focusing in a radial electrostatic field is:

$$(d_{1e} - g_e)(d_{2e} - g_e) = f_e^2 \qquad (11)$$

where g_e and f_e are defined by the equations:

$$g_e = r_e \cot(\sqrt{2}\,\phi)/\sqrt{2}$$
$$f_e = r_e/\sqrt{2} \sin \sqrt{2}\,\phi;$$

d_{1e} is the distance from the apex of the divergent beam to the effective face of the electrostatic field; d_{2e} is the distance from the effective exit face of the electrostatic filed to the point of focus; r_e is the radius of curvature of the particle path in the electrostatic field; and ϕ is the angle through which the ions are deflected.

As an example of this equation consider the symmetrical focal points for electrostatic deflections of 90°, Fig. 3. The focus eq. (11) becomes for this case:

$$d_{1e} = d_{2e} = r_e\,(\cot 127.28° + 1/\sin 127.28°)/\sqrt{2} \qquad (12)$$

or

$$d_{1e} = d_{2e} = 0.35r_e \qquad (13)$$

Thus, the object and image distances for the 90° electrostatic lens are $0.35r_e$ units from the effective ends of the condenser field. A slight variation of this type of electrostatic lens was first reported in mass spectroscopy by Dempster,[18] who used it in connection with his double focusing mass spectrograph.

3. *Velocity Focusing Properties of a Combination of Electrostatic and Magnetic Fields*

By proper combinations of electrostatic and magnetic fields a collimated beam of ions of one mass, but heterogeneous in energy may be focused. In the special case of a direction focusing electrostatic field followed by a direction focusing magnetic field as shown in Fig. 4, the condition for velocity focusing as given by Mattauch and Herzog[19] is:

$$\begin{aligned}[r_e(1 - \cos\sqrt{2}\,\phi) + \sqrt{2}\,(\Delta - d_{1m})\sin\sqrt{2}\,\phi]\\ = \pm[r_m(1 - \cos\theta) + d_{1m}\{\sin\theta + \tan E(1 - \cos\theta)\}] \quad (14)\end{aligned}$$

where E is the angle between the normal to the magnetic field and the incident ion beam. The positive sign is used in this equation when the ion deviations in the two fields are in the same direction, and the negative when in the opposite. The condition for "*double focusing,*" i.e., simultaneous correction for direction and velocity inhomogeneities, is that eqs. (8), (11), and (14) must be satisfied simultaneously. Several machines which accomplish this have been built and are summarized in section III-6.

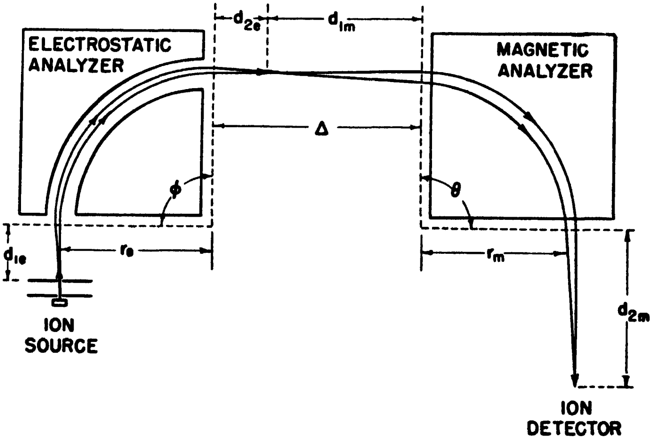


FIG. 4.—Diagram illustrating the double focusing action of a combination of direction focusing electrostatic and direction focusing magnetic fields.

It is apparent from eq. (14) that in the general case, the condition of double focusing can be attained at only one point, i.e., at $r_e = kr_m$. This is not a factor when a double focusing machine is used to focus an ion beam of one mass on a fixed slit. However, when the beams of different mass are focused on a photographic plate to form a mass spectrum it is of prime importance to know how rapidly the machine deviates from the perfect focusing condition.

For more general mathematical details on the focusing conditions the reader is referred to the papers of Bartky and Dempster,[20] Herzog,[15] Mattauch and Herzog,[19] Dempster,[21] Bainbridge and Jordan,[16] Cartan,[22] and Hutter.[23]

4. Mass Dispersion Produced by a Magnetic Analyzer

The approximate distance between the ion beams of two different masses at the point of focus is:

$$D = (r_m/2)(\Delta m/m_0)(1 + f_m/(d_{1m} - g_m)) \tag{15}$$

where D is the distance measured perpendicular to the ion beam between the two masses in centimeters, Δm is the difference in mass of the two masses under consideration, and m_0 is the mass of the focused ion beam.

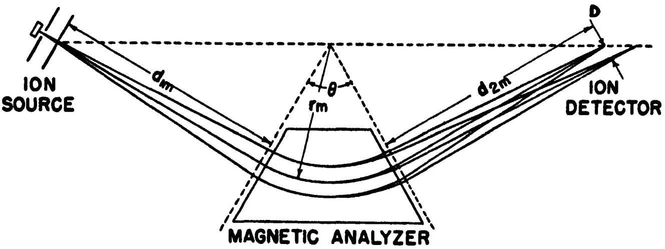


FIG. 5.—Graphical construction of the 60° sector type symmetrical magnetic analyzer, showing the focusing action and the mass dispersion.

This equation is valid for ions entering the analyzer at constant energy. For those entering at constant velocity the dispersion is doubled.

If the analyzer is of the symmetrical type, $d_{1m} = d_{2m}$, substitution of eq. (8) gives:

$$D = r_m \Delta m/m_0 \tag{16}$$

Thus, for example a 60° symmetrical magnetic analyzer operating with a mean radius of curvature of 50 cm. has a mass dispersion of 10 mm. for a 1% mass difference as measured along the locus of the focal points, see Fig. 5.

5. Line Widths Produced by Electrostatic and Magnetic Analyzers

It is impossible to give in a short article, such as this, complete discussions of line widths produced by analyzers. It will be instructive, however, to write down the equation for the image widths produced by symmetrical electrostatic or magnetic analyzers. The equation for the electrostatic analyzer obtained by Stephens[24] is:

$$W_E = (\tfrac{4}{3})(r_e\alpha^2) + S_1 + r_e\Delta V/V \tag{17}$$

where W_E is the width of the ion beam at the image point in centimeters, α is the half angle of divergence of the ion beam emanating from the object point in radians, S_1 is the width of the defining slit at the object point in centimeters, V is the average energy of the ions beam, and ΔV is the spread in energy in the ion beam.

Similarly for the symmetrical magnetic analyzer:

$$W_M = r_m\alpha^2 + S_1 + r_m\Delta V/V \tag{18}$$

where the definitions apply as before, except that they apply to the magnetic analyzer.

It should be noted, that the $r_m\alpha^2$ term of eq. (18) is the term which results when plane faces are used for the magnetic analyzer. By properly shaping the entrance, and/or, exit faces of the magnetic analyzer or by making the magnetic field nonuniform, this term can be eliminated. The radius of curvature of the field entrance and exit faces for the 60 and 90° symmetrical analyzers, which will compensate for the $r_m\alpha^2$ term of eq. (18), has been worked out by Bainbridge. The solution is approximately:

$$r_H = r_{om} \cot^2 \frac{\theta}{2} \left(\cot \frac{\theta}{2} - \alpha\right) \tag{19}$$

where r_H is the radius of curvature of the field boundaries, and r_{om} is the radius of curvature of the median ray.

Several practical considerations come into the use of eqs. (17) and (18).

(1) They assume that the fields are homogeneous.
(2) They assume there is no Rutherford type scattering of the ions due to gases in the analyzer chamber, i.e., the pressure is low.
(3) They assume that there is no dispersion of the ion beam due to electrostatic repulsion among the charged particles.
(4) They assume that there is no deviation of the ion beam due to the presence of surface effects in the analyzers.

From this discussion the important factors, those that influence line widths in the general case, are immediately apparent.

III. APPARATUS

A number of physical arrangements have been used by the different workers in the field of mass spectroscopy to solve different problems. For convenient discussion the machines are divided into four components, see Fig. 5. The source which forms a collimated beam of ions, and the sample system by means of which the sample is introduced into the source, the analyzing system by means of which the ion beams are separated according to their M/e value, and the detecting system by means of which the resolved ion beams are collected and measured.

1. Ion Sources Used in Mass Spectroscopy

Table I summarizes the various types of ion sources which have been used in mass spectroscopy. For actual details of construction the reader is referred to the references listed in Table II.

TABLE I. Ion sources used in mass spectroscopy.

m source type	Approximate spread in ion energy (volts)	First reported by	Special requirements	Best use for
Gaseous discharge	1000	Thomson[25]	Double focusing spectroscope	Packing fractions
Anode ray	1000	Aston[26]	Double focusing spectroscope	No longer used
Hot anode	0.2	Dempster[18]	Work function of element must be low	Isotopic abundance and chemical purity
Hot anode in gas	0.2	Moon and Oliphant[27]	Work function of element must be low	Isotopic abundance
Electron bombardment	0 2-4 0	Dempster[28]		General purpose
Hot spark	1000	Dempster[29]	Double focusing spectroscope	General purpose
Gaseous discharge in magnetic field	100	Wall[30]	Double focusing spectroscope	Leak detection
Secondary ion	5	Smith[31]		Surface phenomena
Arc discharge	10	Koch[32]		Isotope separation

The gaseous discharge type of source has been used in recent years for packing fraction determinations only. Even in this application it is now being supplanted by the electron bombardment source, which is much more flexible, and in which the conditions of ionization are under much better control.

The anode ray source is of historical importance only.

The hot anode source, one arrangement of which is illustrated schematically in Fig. 6, is of considerable use in analyzing elements which have low ionization potentials. With this type of source the material to be analyzed is applied directly to the filaments as a solid. When this filament is heated in vacuum, some of the material is evaporated from the filament as ions. These ions are then accelerated and collimated into an ion beam by the collimating slits of the source.

The ionization produced by the hot anode source depends on the fact that the filament has a higher affinity for electrons than the material applied to it. Specifically, the degree of ionization of the material evaporated from the filament can be written:[33]

$$n^+/n^\circ = e^{F(W-IP)/RT} \tag{20}$$

where n^+/n° is the ratio of positively charged to neutral particles evaporated, F is the Faraday number, W is the work function of the filament material, IP is the ionization potential of the material evaporated from the filaments, R is the gas constant, T is the absolute temperature, and e is the naperian base. This formula applies directly only to the elements, but it is qualitatively correct for compounds. It is apparent from eq. (20) that as the ionization potential decreases or the work function increases, the degree of ionization goes up. For this reason the filament must be made of material with a high work function. Platinum with a work function of 6.2 volts is one of the best. However, for materials of very low vapor pressures this material cannot be used at sufficiently high temperature. For this reason tungsten, with a work function of 4.52 volts, is more generally usable. Below 1800°C. this material may be oxygenated, increasing the work function to a maximum of about 9.2 volts.

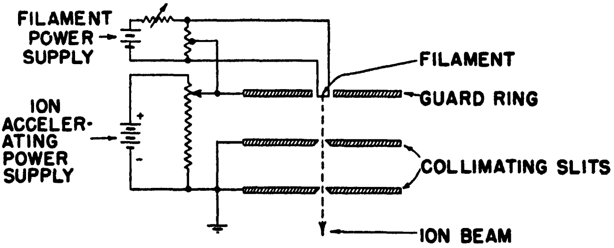


FIG. 6.—Schematic diagram of the hot anode source. With this source the ions are formed by evaporation of ions from the surface of the filament.

The hot anode source is very selective in its ionization. For example, if a mixture of lanthanum and nickel salts is applied to such a filament, only the lanthanum will ionize. Thus, the source is adaptable for detecting certain impurities present in very small amounts. It is especially advantageous since it does not ionize the background gases. The spectrum obtained with this type of source is thus freer of impurities than the more generally used electron bombardment source. One recent variation of the simple hot anode source is the arrangement worked out by Shaw[34] in which he places the sample to be analyzed in a tungsten crucible, which is heated by electron bombardment.

Among the elements which can be analyzed with this type of source are lithium, sodium, aluminum, potassium, calcium, rubidium, strontium, gallium, yttrium, zirconium, ruthenium, rhodium, indium, cesium, barium, lanthanum, cerium, praseodymium, neodymium, element 61, samarium, europium, gadolinium, terbium, dysprosium, holmium, erbium, thulium, ytterbium, lutecium, hafnium, tungsten, and uranium.

It should be noted that there are also a number of elements which evaporate from surfaces as negative ions, however, ionization efficiencies are always very low. The fourth type of source listed in Table I is simply a variation of the above source. With this source the material to be ionized is introduced as a gas rather than placed directly on the filament as a solid. It is very convenient to use when the material to be investigated can be put in the gaseous form, since the problem of sample introduction is greatly simplified.

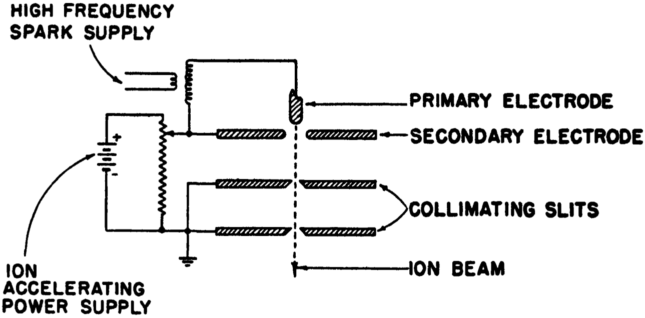


FIG. 7.—Schematic diagram of the hot spark source. Ionization is produced in this source by maintaining a high frequency spark discharge between the two electrodes.

The hot spark ion source was applied to mass spectroscopy by Dempster.[29] A typical electrode arrangement is shown schematically in Fig. 7. With this source an oscillating circuit maintains a spark discharge between a central primary electrode and the edges of a hole in the secondary electrode. The primary electrode is either constructed of the material to be analyzed or it is a tube packed with the material to be analyzed. When the discharge takes place between the electrodes some of the primary electrode is vaporized and ionized. The ions formed are accelerated and collimated into an ion beam by the collimating slits.

This spark source is one of the most universal sources now in use. It will analyze any element that can be put into a solid form, either as the element or as a compound. It is especially valuable for packing fraction work as it produces an abundance of multiply charged ions.

The major disadvantages of the source are that the ion beam is unstable and that the ion beam emanating from the source has a spread in energy of the order of 1000 volts. It thus requires a spectrograph or spectrometer which includes velocity focus or velocity selection in addition to mass resolution.

The electron bombardment source was first used by Dempster.[28] It was further developed to its present highly reliable state by Smythe,[64] Bleakney,[35] Tate and Smith,[36] and Nier.[17,37] Ionization in this source is produced by electron bombardment of gases. In the mass spectrometric application electron ionizing energies of 75–100 volts are used. However, when the source is used for packing fraction work where multiply charged ions are an advantage (see section IV-2) electron energies of 1000 volts are often used. Under these condition heavy mass ion with eight charges are obtainable in sufficient quantity to be usable.

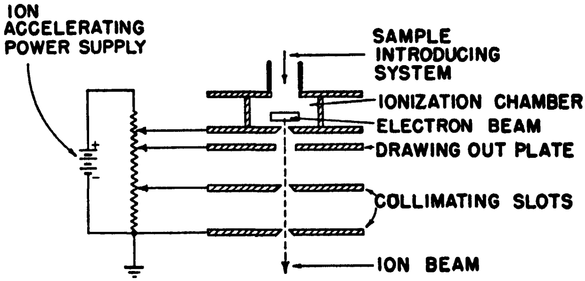


FIG. 8.—Schematic diagram of the electron bombardment source. Ionization of the gas in the ionization chamber is accomplished by bombardment with electrons. The electron beam in the figure is directed normal to the page.

A typical source of this type is illustrated in Fig. 8. In this particular diagram the electron beam is directed across the ionization chamber perpendicular to the page. For most applications the electron beam is held in alignment with a magnetic field. The ionization chamber is constructed with as few and small openings as is compatible with efficient removal of the ions. This gives a high ratio of inside to outside pressure so that source efficiencies are high. In the special case of hydrocarbon analysis the filament which produces the ionizing electron beam must be well isolated from the ionization chamber so that gases cracked by the filament do not interfere with the analysis, and so that the temperature of the ionization region remains constant. These precautions are not necessary in a source used for isotope abundance or packing fractions.

One important point in the operation of this type of source is that the efficiency of the source varies as the potential across the source is varied. This effect is reduced by maintaining the potential between the ionization chamber and the drawing out plate constant, but it is by no means eliminated. This effect is of such magnitude that absolute isotopic abundances determined by varying the ion acceleration voltage are noticeably in error.

Samples are introduced to this source through the hole shown at the top of the source. Materials which have appreciable vapor pressures at room temperature are admitted to the source through a "gas leak." Materials which are nonvolatile at room temperature, but have considerable vapor pressure below 1500°C. are evaporated into the source from a crucible or crucibles places so that a molecular beam of the sample crosses the electron beam. Materials of lower vapor pressure such as platinum, tungsten, etc., are evaporated from filaments similarly placed.

This electron bombardment source is the source most used in modern mass spectroscopy. It has the advantages of extreme stability, and low energy spread. It can thus be used without velocity focusing or selection except when packing fractions are to be measured.

The source utilizing a gaseous discharge in a magnetic field is a high current source based on the principle used in the Phillips ion gauge.[38] It is in actuality, simply a variation of the gas discharge originally used by Thomson.[25] The source has not been very widely used because of its inherent instability. It gives ions of energy ranging to ±100 volts from the average depending on the particular design. Thus, for most applications velocity selection or velocity focusing must be used in addition to mass resolution. It is also of interest in that it gives ions due to ion bombardment of the discharge chamber surfaces. Thus, for example, if the discharge chamber is made of copper the ion beam emanating from the source will include copper ions.

The secondary ion source is of interest in the study of surfaces, or in the analysis of solid samples. The method of operation is to direct an ion beam of some inert gas onto a solid sample; when this ion beam strikes the sample, ions characteristic of that surface are formed.

The arc source is of interest mainly in forming an intense ion beam for separation of measurable quantities of isotopes. It has not been used for other purposes.

2. Sample Handling Systems

All the discussion of sources has assumed that the intensity of the ion beams formed by the source are always characteristic of the sample. This is by no means true unless special precautions are taken in the introduction of the sample.

It is impossible in a short article to give a complete discussion of mass discrimination effects. However, an outline of some of the problems involved in the introduction of gaseous samples to the mass spectrometer will serve to illustrate the types of problems encountered.

Knudsen[39,40] and later Smoluchowski[41] derived the equations of flow for gases at low pressures through tubes and orifices. The molecular flow of gas through cylindrical tubes can be represented approximately by the equation:

$$Q_m = 3800(D_T^3/L)(T/M)^{\frac{1}{2}}(P_1 - P_2) \text{ dyne-cm./second} \tag{21}$$

and the flow through circular openings by:

$$Q_m = 2860D_T^2(T/M)^{\frac{1}{2}}(P_1 - P_2) \text{ dyne-cm./second} \tag{22}$$

where Q_m is the molecular gas flow through opening in dyne-cm./second, D_T is the diameter of opening in centimeters, L is the length of tube in centimeters, M is the molecular weight on the scale $0 = 16$, P_1 is the pressure on the high pressure side of opening in dynes/cm.^{2}, and P_2 is the pressure on the discharge side of the opening in dynes/cm.^{2}

Theoretically these equations hold as long as the mean free path (λ) of molecules is large compared with the diameter of the opening ($\lambda \gg D_T$). Experimentally it has been found to hold well, providing the mean free path is at least twenty times the diameter of the opening.

For high pressure ($\lambda \ll D_T$) the gas flow is viscous and follows Poiseuille's law:[42]

$$Q_v = (\pi D_T^4/256\eta L)(P_1 + P_2)(P_1 - P_2) \tag{23}$$

where Q_v is the viscous gas flow through the opening in dyne-cm./second, and η is the viscosity in poises.

To determine the limits for nonseparative flow it is convenient to evaluate the diffusive mixing term. The rate of flow of one component is given by the equation:

$$Q_1 = \frac{\pi D_T^2}{4}\left(vn_1 - D\frac{\partial n_1}{\partial L}\right) \tag{24}$$

where Q_1 is the net flow of component, D is the diffusion coefficient, n_1 is the number of molecules/cc., and v is the drift velocity of the gas through the tube.

Obviously, if the term depending on v is much larger than the term dependent on D, the flow through the tube will be nonseparative, i.e., there is no fractionation in the sample reservoir due to back diffusion. If, however, the term depending on D is comparable to that depending on v, back diffusion will take place through the tube and indeterminable errors will be introduced. These equations are sufficient for determining the characteristics of a gaseous sample introduction system. It now remains to set up the requirements for an ideal system and show how closely a practical system meets these requirements.

The requirements for ideal operation of a leak for introducing gaseous samples into a mass spectroscope operating with an electron bombardment source are:

(1) The composition of the gas mixture in the ionization region of the source should be identical with that of the sample.
(2) The concentration of the mixture being analyzed should not change with time.
(3) In a gas mixture, a change in the amount of one substance should not affect the peak height due to the others, i.e., all mixture peaks should be linear superpositions of the individual intensities.
(4) The gas flow should remain essentially constant during the analysis.

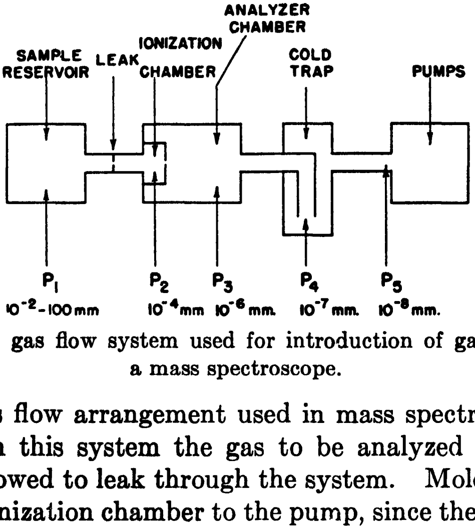


FIG. 9.—Schematic gas flow system used for introduction of gaseous samples into a mass spectroscope.

A typical gas flow arrangement used in mass spectroscopy is shown in Fig. 9. With this system the gas to be analyzed is placed in the reservoir and allowed to leak through the system. Molecular flow takes place from the ionization chamber to the pump, since the mean free paths are large compared with the openings, that is, pressures are low throughout this region.

Consider now the molecular flow type of leak, that is, one whose diameter is less than one twentieth of a mean free path. In equilibrium the gas flow into the ionization region Q_{12} must equal the flow out Q_{23}. Since both flows are molecular we have from eq. (22):

$$2860D_{12}^2(T/M)^{\frac{1}{2}}(P_1 - P_2) = 2860D_{23}^2(T/M)^{\frac{1}{2}}(P_2 - P_3) \tag{25}$$

Making the approximation that $P_1 \gg P_2$ and $P_2 \gg P_3$:

$$P_2 = kP_1 \tag{26}$$

where k is a constant determined by the dimension only. This shows that the composition of the mixture in the ionization chamber is identical to that in the reservoir. Thus, this leak meets the first requirement of leaks.

The second requirement, however, is not met by the molecular flow leak, since the composition of the material in the reservoir changes with time. This change in concentration for a binary system is represented by the equation Honig[43] gives in his discussion of sample introductions systems:

$$(P_1/P_0)_a = (P_1/P_0)_b^{(M_b/M_a)^{\frac{1}{2}}} \tag{27}$$

where $(P_1/P_0)_a$ is the ratio of final pressure to initial pressure in the reservoir for component a, $(P_1/P_0)_b$ is the ratio of final to initial pressure in the reservoir for component b, M_a is the molecular weight of component a, and M_b is the molecular weight of component b. For example, if a sample containing 50% benzene (C_6H_6) and 50% hydrogen (H_2) is allowed to pass through a leak until the benzene pressure in the reservoir has dropped by 6%, the composition will change to 42% hydrogen and 58% benzene. Obviously under these conditions accurate analyses are impossible. The method of surmounting this difficulty is to use a reservoir of such size, that the change in composition is less than the accuracy to which the sample is to be measured.

The molecular flow leak does satisfy the third requirement in that the peaks are superposable. This fact results very simply from analysis of the equation of flow, and is apparent also from the fact, that in molecular flow every molecule acts as though no others were present. The fourth requirement can be met only approximately, i.e., by using a reservoir, and hence a sample of sufficient size so that the change in pressure during the time of analysis is negligible, or by using a constant pressure device on the sample reservoir.

There is little doubt that the molecular flow leak is the best where gas analyses are involved, for example, the relative abundances of benzene and hydrogen. It is not necessarily the best when routine isotopic comparisons are to be made.

The second possible leak type is one that works on mass flow principles. The condition worked out by Nier[12] for this case is that the flow through the tube is sufficiently fast, and the pressure sufficiently high, so that back diffusion does not serve to mix that part of the gas that is fractionated by molecular flow with that coming from the sample. That is, in eq. (24) the velocity term is very large compared to the diffusive mixing term. Even though the flow through the leak will be governed by mass flow principles, the flow throughout the rest of the instrument is molecular. Thus, to a fair approximation the equation for the equilibrium flow ($Q_{12} = Q_{23}$) is:

$$(\pi D_{12}^4/256\eta L)(P_1 + P_2)(P_1 - P_2) = 2860 D_{23}^2 (T/M)^{\frac{1}{2}} (P_2 - P_3) \tag{28}$$

Again making the approximation that $P_1 \gg P_2$ and $P_2 \gg P_3$:

$$P_2 = [\pi D_{12}{}^4/(256)(2860)D_{23}{}^2\eta L](M/T)^{\frac{1}{2}}P_1{}^2 \tag{29}$$

or the equation may be written:

$$P_2 = \frac{K(M)^{\frac{1}{2}}}{\eta(T)^{\frac{1}{2}}}P_1{}^2 \tag{30}$$

where K is a constant determined by the dimensions only. This shows that the first requirement of leaks is not satisfied, because the pressure of the sample in the source is not independent of the molecular weight of the material in the sample reservoir. In binary problems the relations are quite unpredictable. The second requirement of leaks, however, is satisfied by this leak because the composition of the material in the reservoir does not change during analysis. The third requirement is not satisfied. For example, 1% of benzene added to a helium sample will change the helium peak height by about 50%. There is, however, one factor which makes the leak important. It is the fact that the ratio of any two peaks is characteristic of the relative composition only. The fourth requirement can again be met by using large samples and containers. In this one respect, the two leaks are similar.

Thus, the mass flow leak system can be used for normal isotopic abundance only if the appropriate square root of the mass correction is applied. However, it is probably the better leak for use in routine isotopic comparisons, since in this case all that is required is that the fractionation holds constant, while the fact that no isotopic composition change takes place in the reservoir is more important. For example from eq. (27) it is seen that if 10% of a standard sample of HD is allowed to leak through a molecular flow leak the ratio of the isotopes in the sample would have changed in concentration by 2.4%. This would require discarding the standard. However, if a mass flow leak is used no such errors are introduced.

It is possible to make variable leaks of the mass flow type. One such leak has been described by Nier, Ney, and Inghram,[44] which, if operated at the correct back pressure and leak size, satisfies mass flow requirements.

In the case of the hot anode source, where the gas is directed at a heated filament as a molecular beam, the requirements are just the opposite of the case just discussed. In this case, the molecular flow leak gives a fractionation factor equal to the square root of the mass ratio, while the mass flow leak gives no discrimination.

If the sample is introduced into a crucible and heated to give a molecular beam, which is ionized by electron bombardment, two corrections must be considered. The first correction arises from the fact that the rate of evaporation is different for molecules of different mass, i.e., from equiporation of energy it may be shown that the lighter mass molecule moves faster, and hence has a higher probability of escape than the slower. The evaporation rate is inversely proportional to the square root of the ratio of the masses.[45] Thus, the composition of the gas being bombarded by electrons is different from that in the crucible, and the correction must be applied. It should be noted that this simple factor does not hold for light elements such as lithium, nitrogen, etc. There is a complicating factor in this simple evaporation assumption. This is that the fractionation will take place only if there is perfect mixing of molecules at the sample face. If there is no mixing this fractionation cannot take place; there is a mass motion from the sample and hence no fractionation. It may be assumed that perfect mixing takes place in the volatilization of a liquid and that none takes place in the volatilization of a solid. This is only approximately true since there is some mixing in solids at high temperatures. The second correction to be applied to the crucible evaporated samples comes in from the fact that the molecules traverse the ionization region at different speeds and are condensed on the walls of the chamber at the first collision. This introduces another square root of the mass factor, which either counterbalances the factor obtained from the evaporation effect of liquids, or introduces a correction which must be applied to the nonfractionating evaporation of a solid.

The above factors include the most important types of discrimination, which occur in the introduction of samples to the mass spectrometer. Obviously, only by the careful evaluation of the type of answer that is desired can the proper arrangement of leaks be made.

3. Mass Analyzers

The various types of analyzers, which may be used for separation of an ion beam into its various mass components, can best be illustrated by referring to the various mass resolving instruments which have been constructed. The most important and instructive of these instruments are summarized in Table II.

The first mass analyzer was designed and constructed by J. J. Thomson.[25] With his analyzer the arrangement of fields was such that the ion beams of different mass components were separated into different parabolas. The principle significance of his method is historical. There is no focusing with this type of machine so the lines are not of sufficient sharpness to be of practical value in modern mass spectroscopy.

The first use of the focusing action of magnetic fields was reported by Dempster in 1918.[18] The arrangement he used was the 180° direction

TABLE II. Summary of important mass spectroscopes which have been described in scientific periodicals.

Workers	References	Year	Type of machine	Magnetic angle	Electrostatic angle	Primary purpose
Thomson	1	1913	Parabola	Simultaneous and variable		Isotope existence
Dempster	18, 28	1918	Direction focusing	180°	None	Isotope existence and abundance
Aston I	46	1919	Velocity focusing	Variable (eq. 28)	5° 42′	Isotope existence Packing fractions
Costa	47	1925	Velocity focusing	Variable (eq. 28)		Packing fractions
Aston II	48	1927	Velocity focusing	Variable (eq. 28)	9° 33′	Packing fraction
Bleakney	49	1929	Velocity selection Direction focusing	180°	0° Wien filter	Ionization phenomena
Bainbridge	50	1930	Velocity selection Direction focusing	180°	0° Wien filter	Packing fraction
Smyth and Mattauch	51, 52, 53	1932	Oscillating electric fields	None	90°	Isotope abundance
Bleakney	35	1932	Direction focusing	180°	None	Isotope abundance
Oliphant, Shire, and Crowther	54	1934	Velocity selector	Wien filter		Isotope separation
Tate and Smith	36	1934	Direction focusing	180°	None	Isotope abundance
Bondy, Johannsen, and Popper	55	1934	Double focusing	Simultaneous	127°	Isotope abundance
Smythe, Rumbough, and West	56, 57, 58	1934	Special direction focusing	Variable	None	Isotope separation
Dempster	29	1935	Double focusing	180°	90°	Packing fraction
Bainbridge and Jordan	16	1936	Double focusing	60°	127° 17′	Packing fraction
Sampson and Bleakney	59	1936	Direction focusing	180°	None	Isotope abundance
Aston III	60	1937	Velocity focusing	Variable (eq. 28)	14° 20′	Packing fraction
Mattauch	61	1936	Double focusing	90°	31° 50′	Packing fraction
Nier	62	1936	Direction focusing	180°	None	Isotope abundance

TABLE II. Summary of important mass spectroscopes which have been described in scientific periodicals.—(*Continued*)

Workers	References	Year	Type of machine	Magnetic angle	Electrostatic angle	Primary purpose
Bleakney and Hipple	63	1938	Double focusing	Simultaneous	180 + 360°	Ionization phenomena
Nier	17	1940	Direction focusing	60°	None	Isotope abundance
Jordan	65	1940	Double focusing	60°	Velocity filter	Packing fraction
Hoover and Washburn	5, 7, 8	1940	Direction focusing	180°	None	Gas analysis
Straus	66	1941	Double focusing	60°	90°	Isotope abundance
Brown, Mitchell, and Fowler	67	1941	Direction focusing	60°	None	Isotope abundance
Hipple	6, 68	1942	Direction focusing	90°	None	Gas analysis
Coggeshall and Jordan	69	1943	Direction focusing	60°	None	Isotope abundance
Nier, Inghram, and Stevens	70	1943	Direction focusing	60°	None	Hydrogen isotope analysis
Taylor	71	1944	Direction focusing	60°	None	Isotope abundance
Thomas, Williams, and Hipple	72	1946	Direction focusing	180°	None	Isotope abundance
Thode, Graham, and Ziegler	73	1945	Direction focusing	180°	None	Isotope abundance
Stephens	74	1946	Pulse type velocity selector	None	None	Gas analysis
Forrester and Whalley	75	1946	Direction focusing	60°	None	Panoramic scanning
Nier, Stevens, Hustrulid, and Abbot	11	1947	Direction focusing	60°	None	Leak detection
Nier (others)	12	1947	Direction focusing	60°	None	Continuous plant analysis
Shaw and Rall	76	1947	Double focusing	90°	31° 50′	Solid analysis
Nier	37	1947	Direction focusing	60°	None	Isotope abundance
Siri	77	1947	Direction focusing	180°	None	Gas analysis for medical purposes

focusing magnetic analyzer to be discussed in more detail later in this section.

The second mass spectrograph, which produced a focused spectrum of the mass components, was the velocity focusing spectrograph of Aston.[46] This spectrograph is illustrated schematically in Fig. 10. The condition for velocity focus with this apparatus is:

$$\theta = \frac{2\phi(d_1 + d_2)}{d_2} \tag{31}$$

With the position of the detector photographic plate shown, the velocity focus equation is met for all parts of the plate. With this apparatus there is no direction focusing. This means that the ion beam must be defined very closely with the collimating slits S_1 and S_2. This is a limitation on the Aston type of machine, which was later overcome by other workers with the double focusing technique. As a result of this difficulty this type of machine, which has a glorious past, is no longer used.

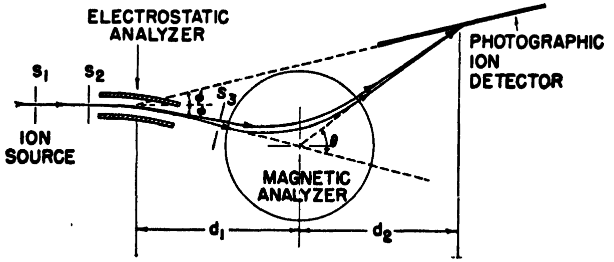


FIG. 10.—Aston's velocity focusing mass spectrograph. With this machine there is little direction focusing so that the ion beam must be defined very closely by the slits S_1 and S_2.

Modern mass spectroscopes can be classified into one of four groups. The first is the direction focusing magnetic analyzer; the second is the velocity selection-direction focusing analyzer; the third is the double focusing analyzer; and the fourth is the velocity selector analyzer. These groups are sufficiently distinct to merit separate discussions.

4. Direction Focusing Magnetic Analyzers

Direction focusing magnetic analyzers are the simplest mass analyzers used in modern mass spectroscopy. Their value is well illustrated by noting the large number of this type of machine listed in Table II. The analyzer consists of a single homogeneous magnetic field with the source and collector arranged according to the direction focus eq. (8). Only those sources which produce ion beams quite homogeneous in energy, see Table I, can be used with these analyzers.

One of the simplest types of direction focusing magnetic analyzers, is that in which the source slit is at the entrance to the magnetic field and the collector at the exit, i.e., $d_{1m} = d_{2m} = 0$. If these values are substituted in eq. (8) we have:

$$g_m = \pm f_m \tag{32}$$

or

$$\theta = n\pi \tag{33}$$

where n is any integer. Thus, direction focusing will take place if the ion beam deviates through 180°. This is exactly the case first used by Dempster[18] in applying direct focusing to ion optics. His apparatus is illustrated schematically in Fig. 11. Since that time a large number of machines have been constructed which utilized this focusing angle, see Table II.

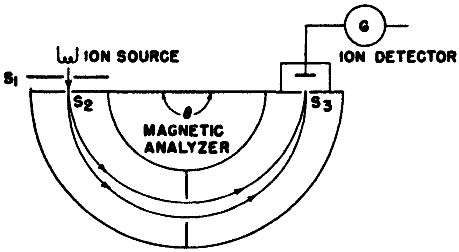


FIG. 11.—Schematic diagram of the 180° direction focusing mass analyzer used by Dempster in 1918.

More recently, the sector shaped magnetic analyzer has come into use. The most common case is the 60° symmetrical analyzer of Nier,[17,37] illustrated in Fig. 5, and the 90° symmetrical analyzer of Hipple.[6] The condition for focus in the 60° case was computed in section II-3. The condition for the 90° case is computed by substituting 90° for θ in eq. (8). Thus, the symmetrical object and image lengths are:

$$d_{1m} = d_{2m} = r_m \tag{34}$$

The dispersion along the locus of the line of focus for the 180°, 90°, and 60° instruments are, for identical radii of curvature, in the ratio of 1:1.41:2.0, respectively, as is proved by analysis of eq. (15). However, the resolving power of the three symmetrical machines are identical. In the 180° case, the dispersion varies along the locus of the focal points according to the square of the mass, the 60° case is almost exactly linear, and the 90° case falls intermediate.

From this discussion it is obvious that there is no fundamental difference between any of the direction focusing magnetic analyzers. Any one of them, if proper care is taken in the details of design and operation, will do equally well any of the jobs required of a direction focusing mass analyzer.

5. *Velocity Selection—Direction Focusing Analyzers*

The requirement of the simple direction focus machine, that the energy of the ion beam entering the magnetic analyzer be quite homogeneous, is removed by the addition of a velocity selector. Such a velocity selector picks out a homogeneous beam of ions which is then projected into a direction focusing magnetic analyzer.

It must be emphasized at this point that the types of analyzers discussed under these separate headings are distinct, and that each is best for particular types of jobs. The succeeding sections in this discussion should not be taken to mean improvements over earlier types. For example, the simple direction focus machine can be used to measure isotopic abundances but cannot be used for packing fractions, while the velocity selection-direction focusing analyzer is well suited for measuring packing fractions, but is completely unusable for accurate measurement of isotopic abundances.

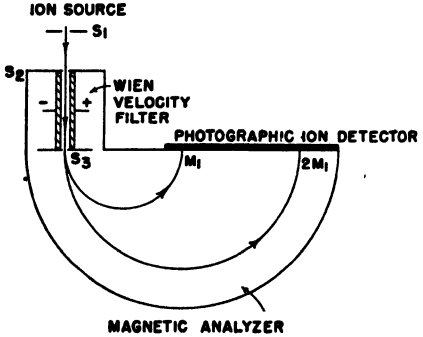


FIG. 12.—Schematic representation of the velocity-selector-direction focusing mass analyzer used by Bainbridge. The velocity selector is the crossed field Wien filter.

The addition of velocity filters was undertaken by Bleakney in 1929[49] and by Bainbridge in 1930.[50] The arrangement is shown schematically in Fig. 12. The velocity selector is the so-called Wien filter, which directs a beam of ion, homogeneous in velocity, into the 180° direction focusing mass analyzer. The filter consists of crossed electrostatic and magnetic fields both at right angles to the direction of propagation of the central ion beam. Slits are placed so that only those ions which pass through these fields undeflected are utilized. For this trajectory, the forces on the moving ion resulting from interaction with the electrostatic and magnetic fields must be equal. The magnetic force is:

$$F_m = Be(v/c) \tag{35}$$

the electrostatic force is:

$$F_e = eE \tag{36}$$

equating:

$$v = c(E/B) \tag{37}$$

Thus, this arrangement of fields is velocity sensitive and only those ions having the velocity given by this equation get through the filter and into the direction focusing magnetic analyzer.

An improved machine of this type, which also attained double focusing, see next section, was designed and constructed by Jordan[65] at the University of Illinois in 1940. Unfortunately a complete description of the apparatus has never been published. It is claimed to have a resolving power of one part in 30,000 and an accuracy in the measurement of atomic mass of 1 part/million.

The major disadvantage of this class of machine is that it allows only a very narrow range of masses to be studied or, if it uses an ion source of large energy spread, it has low efficiency.

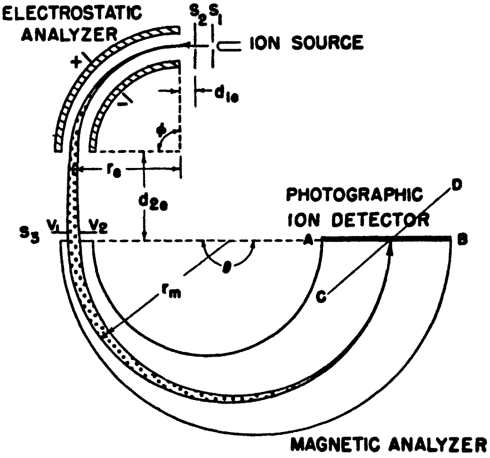


FIG. 13.—Diagram of the Dempster double focusing mass spectrograph. The trajectory of an ion beam inhomogeneous in direction and energy, but of a single mass is shown. This machine is in exact focus at only one position, $r_e = 0.848 r_m$.

6. *Double Focusing Analyzers*

The third major type of mass analyzer is the so-called double focusing instrument. This instrument refocuses beams inhomogeneous both in velocity and direction. Instruments of this type were under development in three independent laboratories simultaneously. The first published was the machine of Dempster[29] Fig. 13. It consists of a 90° direction focusing electrostatic field with $d_{1e} = 1$ cm., $d_{2e} = 5.66$ cm., and $r_e = 8.48$ cm.; and a 180° magnetic analyzer $d_{1m} = d_{2m} = 0$ of mean radius 10 cm.

Since the central ion beam in this machine enters and leaves the analyzer fields at right angles the performance of this machine can be explained by direct substitution in eqs. (8), (11), and (14).

If we start by choosing the values $r_e = 8.48$ cm., $d_{1e} = 1$ cm., and $\phi = 90°$ then substitution in eq. (11) gives $d_{2e} = 5.66$ cm. Thus, the image due to the electrostatic analyzer will be at a perpendicular distance of 5.66 cm. from the end of the electrostatic analyzer. Now using this as the object point for the magnetic field and choosing $d_{1m} = 0$, and $0 = 180°$, substitution in eq. (8) gives $d_{2m} = 0$. Thus, a second image is formed at the exit to the magnetic field. This machine is, therefore, in focus twice from the direction point of view, i.e., at S_3 and at the ion detection.

In satisfying the direction focusing conditions for the above combination of fields no restriction has been placed upon r_m. Thus, the machine is direction focusing for any value of r_m. However, the requirement of double focusing with this particular arrangement of fields restricts r_m to a definite value. Substituting the direction focusing conditions, i.e., $r_e = 8.48$ cm., $\phi = 90°$, $= 5.66$ cm. (see Fig. 4), $d_{1m} = 0$, $\theta = 180°$, and $E = 0°$, in eq. (14) gives $r_m = 10$ cm. Thus, the Dempster 90°–180° spectrograph satisfies both the direction focusing, and velocity focusing conditions at $r_m = 10$ cm., and hence is double focusing.

From this explanation, it is seen that almost any combination of fields can be made to give double focusing by simply varying any two or more parameters simultaneously, and that all are simply particular solutions of the general double focusing conditions.

One simple relation should be pointed out. It is that whenever any combination of symmetrical direction focusing fields is used, no matter what the angles of deflection, that if $r_m = r_e$ then eq. (14) is automatically satisfied.

When a double focusing mass spectrometer is constructed where the ratio of the radii in the two fields is fixed, the above discussion is sufficient. However, if a mass spectrograph is constructed in which a complete spectrum is recorded on a photographic plate r_m is not fixed and the condition for velocity focus is met at only one position on the plate. In the Dempster type spectrograph, calculation shows that the locus of the velocity focus points lie along the line *CD* which is at an angle of approximately 45° to the line *AB*, the locus of the direction focusing points. It turns out that if a slit 2.0 mm. wide is placed at S_3 the line width 1 cm. away from the point of optimum focus is 0.1 mm. Hence, the double focusing condition is good over a usable range.

The second double focusing mass spectrograph was reported by Bainbridge and Jordan in 1936.[16] This machine, illustration in Fig. 14, combined a 127° 17′ direction focusing electrostatic analyzer with a 60° direction focusing magnetic analyzer to give double focusing. Solution of eq. (11) shows that the only solution for an electrostatic direction focusing angle of 127° 17′ is that $d_{1e} = d_{2e} = 0$. The 60° magnetic analyzer is the analyzer previously discussed in section II-3 with $d_{1m} = d_{2m} = 1.732\ r_m$. Substitution of these values in eq. (14) gives as the final condition for double focusing $r_e = r_m$. Again double focusing is accomplished at only one point, i.e., $r_e = r_m$. Calculation of the angle between the locus of the direction focusing points and the locus of the velocity focusing points gives 5°. Thus, the range over which approximate double focus is obtained is considerably wider than in the Dempster machine. Two other observations should be made in comparing these two machines. The first is that the 127–60° machine has twice the dispersion as measured along the locus of the direction focusing points as the 90–180° machines of the same radius, but that it requires only one-third as much field area. The second is that is has an approximately linear mass scale while the 180° machine has one which varies as the square root of the mass. Thus, for equivalent radii of curvature and detector plate size a wider mass range is covered by the 180° machine.

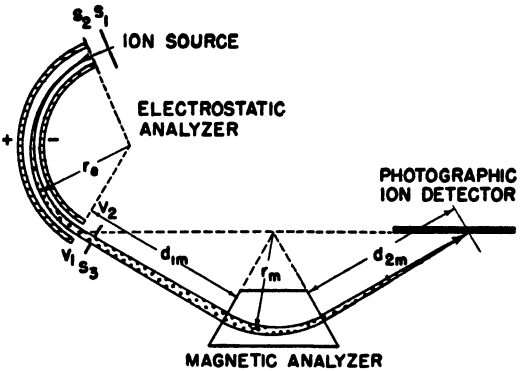


FIG. 14.—Diagram of the Bainbridge-Jordan double focusing mass spectrograph. The trajectory of an ion beam inhomogeneous in direction and energy, but homogeneous in mass is shown. This machine is in exact focus at only one position, $r_e = r_m$.

The third double focusing mass spectrograph to be reported was the 31° 50′ electrostatic field 90° magnetic field machine reported by Mattauch[61] in 1936 and shown in Fig. 15. This machine is different from the two previously described in that the two separate fields are combined to give direction focusing. By solution of eq. (11) it will be found that if the object distance d_{1e} is $r_e/\sqrt{2}$ then the image formed by the electrostatic field is formed at infinity. In other words the ion beam coming from the electrostatic field is parallel. Thus, the beam that enters the magnetic field acts as though it were coming from a source at infinity, $d_{1m} = \infty$. Substitution of this value in eq. (8) shows that if $d_{2m} = 0$

then $\theta = 90°$. Hence, direction focusing is accomplished only by the combination of the two fields. It turns out by analysis of eq. (14) that these restrictions are sufficient to also give velocity focusing.[19] Note that no restrictions have been placed on r_m or Δ, thus the loci of the velocity and direction focusing positions fall along the same line, and the machine is double focusing for all masses simultaneously, and for any value of Δ. Since r_m is independent of r_e it is important to specify some condition to determine what order of magnitude r_e should be. Mattauch showed that the theoretical resolving power is:

$$\Delta m/m = 2S_2/r_e \tag{38}$$

where Δ_m is the least distinguishable mass difference at mass m, and S_2 is the source slit width. Thus, r_e for this machine should be made sufficiently large to give the desired resolution. Again as in the 90–180° case the dispersion as measured along the locus of the focal points varies as the square root of the mass and contrary to the 90–180° and 127–60° cases there is no minimum in the line width because there is no r_m of optimum focus. In the Mattauch case line widths vary directly as the square root of the mass.

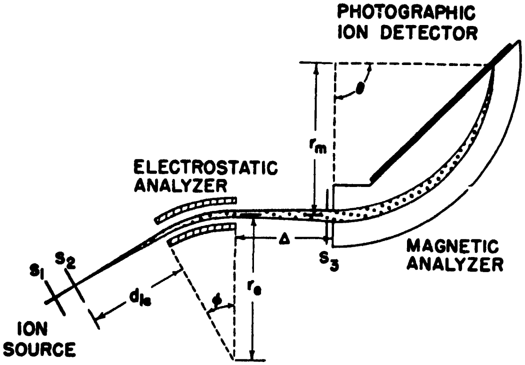


FIG. 15.—Diagram of the Mattauch double focusing mass spectrograph. The trajectory of an ion beam inhomogeneous in direction and energy, but homogeneous in mass is shown. This machine is in focus for all values of r_e, r_m and Δ.

This variation in line width along the plate is of extreme importance in the determination of line strengths by the photographic method since the line widths must be included in the intensity calculations. The effect is not simple in the Dempster or Bainbridge machines since it increases in both directions from the position of optimum focus.

Two other double focusing mass spectroscopes have been constructed, which use simultaneous fields. The first is the double focusing mass spectrometer constructed by Bondy, Johannsen, and Popper[55] according to the theory developed earlier by Bartky and Dempster.[20] This machine is based on the fact that a radial electrostatic field superimposed at right angles to a magnetic field gives double focusing at an angle of 127° 17′ if the force due to the magnetic field is twice that due to the electrostatic field. No recent important work has been done with this type of machine since it suffers from the difficulty of requiring large magnets. It can, however, be used for the leak detection mass spectrometer application.[30]

The second simultaneous field double focusing mass spectrograph is the trochoidal path machine developed by Bleakney and Hipple.[61] This machine gives theoretically better focusing than any of the previously discussed machines. It uses a linear electrostatic field at right angles to the magnetic field. However, it also requires very large homogeneous fields so that it probably will never find wide application.

7. *Velocity Selection Analyzers*

There is only one type of mass analyzer which does not necessarily require a magnetic field. The simple theory is as follows:

Suppose an ion source is used in which the mean energy spread in the ion beam is 0.2 volt, and the total energy is 1000 volts. The velocity of the ion groups will be given by the equation:

$$v = (2eV/m)^{\frac{1}{2}} \tag{39}$$

where v is the velocity in cm./second, V is the voltage through which the ion is accelerated in e.s.u., e is the charge on the ion in e.s.u. and m is the mass of the particle in grams.

To consider a specific example, consider the velocities of the two isotopes of lithum, mass six and mass seven, as emitted from a hot anode source operating at 1000 volts. The velocities of the two isotopes v_6 and v_7 are then:

$$v_6 = 1.825 \times 10^7 \pm 0.002 \text{ cm./second}$$
$$v_7 = 1.691 \times 10^7 \pm 0.002 \text{ cm./second}$$

Thus, the velocities in the ion beam fall into two distinct groups and can be resolved by a properly designed velocity filter.

The first machine to use this principle alone was the Wien filter used by Oliphant, Shire, and Crowther.[54] This filter was, however, used earlier by Bleakney[49] and by Bainbridge[50] in conjunction with a direction focusing magnetic field. The operation of this filter was described in section III-5. There is no focusing action with this type of spectrometer so it can be used only for light masses.

The second machine of this type was suggested by Smythe in 1926 and constructed by Smythe and Mattauch in 1932. This machine operates independently of any magnetic field. The beam passes successively between two condensers of length S_1 separated by a distance S_2. The same alternating field is applied to both condensers. If the frequency of the signal applied to the condensers is $\omega = 2\pi n(v/S_1)$, where n is any integer, the particle suffers as many upward as downward thrusts and the beam leaves the condenser parallel to the incident beam, but displaced by an amount determined by the phase of the signal when the ion enters the electrostatic field. If now the distance of the second condenser is such that the beam travels the distance $S_1 + S_2$ in an odd half cycle, the displacement given by the first condenser will be exactly compensated by the second condenser and the beam of velocity v passes through the combination undeflected; while the others are deflected away from the normal. The machine thus resolves the ions by discriminating between the ion velocities. In practice the machine has a number of ghost images which makes its results difficult to interpret. These effects have been discussed by Hintersberger and Mattauch.[53]

Very recently a machine of this general type has been designed by Stephens.[74] In this machine microsecond pulses of ions are used. The pulsed beam is allowed to pass down a tube until the ions of different mass are divided into distinct velocity groups. Detection is then accomplished by recording the pulses to the final collector with an oscilloscope.

8. Ion Detection Systems

It is the ion detector which determines whether any particular analyzer is a mass spectrometer or a mass spectrograph. Fig. 16 illustrates the type of records obtained with a mass spectrometer, and spectrum (a) of Fig. 17 is the record for the same element obtained with a mass spectrograph. The element used by Inghram in obtaining these particular mass spectra is neodymium. The importance of each type of recording system is immediately apparent from these figures. The type of mass spectrum shown in Fig. 16 is best for measuring the intensities of the various ion beams while the type in Fig. 17 is best for locating accurately the position on the mass scale.

The various types of detector systems which have been used to detect ion beams in mass spectroscopy are summarized in Table III. For more complete details the reader is referred to the machine papers listed in Table II.

One point should be made in connection with the shape of the peaks shown in Fig. 16. In the ideal case the top of the peaks should be flat. When this is true possible variations in ion beam shape due to different mutual repulsions for ion beams of different density or other such effects introduce no errors. The condition for obtaining flat peaks is that the width of the collector slit, for example S_3 of Fig. 9, be greater than the width of the ion beam as defined by eqs. (17) and (18). A second consideration comes in that determines the maximum slit width which can be used. This is the fact that if the collector slit width is equal to the dis-

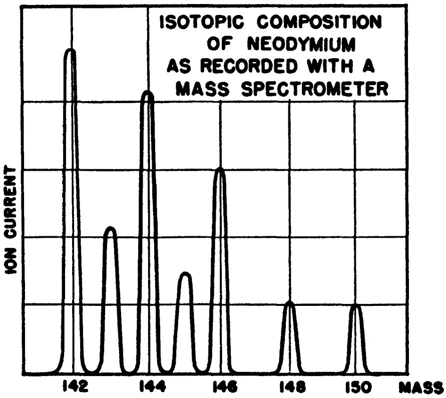


FIG 16—The isotopes of neodymium recorded with a recording mass spectrometer. The ion source used for this particular record was the hot anode source.

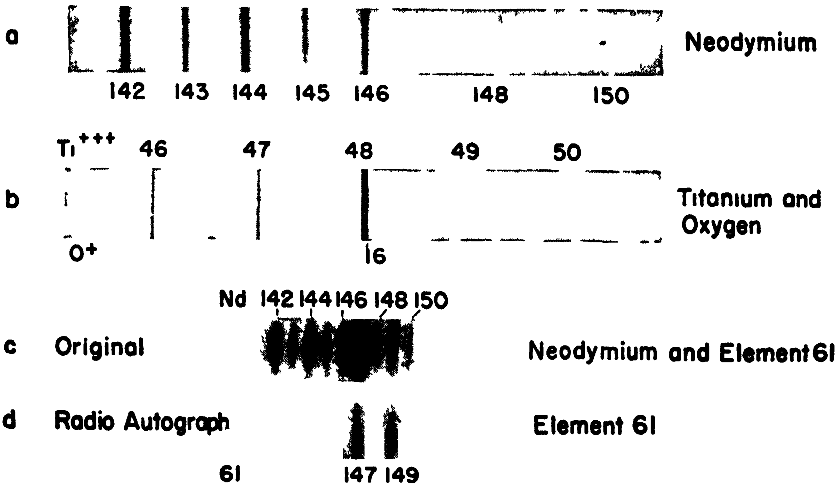


FIG 17—Spectra obtained with a mass spectrograph (a and b) and by disintegration of active deposits on photographic plates (c and d)

tance between two adjacent peaks, no resolution of the peaks is possible. The slit width used in obtaining the curve shown is such that at maximum intensity more than 98% of the total beam is collected.

The first two methods listed in Table III are of historical importance only. They depend on the fact that a rapidly moving beam of charged ions produces a visible fluorescence.

The third method, the technique of allowing the ion beam to strike a photographic plate, is the method of ion detection now used almost universally in mass spectrography. In the past special photographic emulsions have been used for this purpose, among these are Schuman and Ilford Q plates. More recently, the Eastman III-0 ultraviolet sensitive vacuum spectrograph plate has also come into general use. This plate is more rugged than the special Schuman plates and works very satisfactorily with ion energies of greater than 6000 volts. As a typical example, 10^9 ions of 10,000 volts energy and atomic mass 100 striking a square millimeter of area produces a good developable image. It must be cautioned that the image density decreases strongly with an increase in the mass of the ion studied and vise versa. One advantage of the photographic recording is that it integrates the ion beam so that requirements on ion sources are less strigent. Aston[87] has discussed the application of the photographic method in great detail.

TABLE III. Ion detector used in mass spectroscopy.

Original workers	Reference	Type of detector	Primary use
Thomson............	78	Fluorescence	Historical
Dechend and Hammer	79	Photo of fluorescence Direct photo on velox paper	Historical
Thomson............	80	Direct photo on plate	Integrating packing fraction
Thomson............	81	Quadrant electrometer	Intensity measurement
Metcalf and Thomson	82	Electrometer tubes	Intensity measurement
Smith, Lozier, Smith, and Bleakney	83	Electrometer tube and automatic recording	Intensity measurement
Cohen...............	84	Electron multiplier	Intensity measurement
Nier, Ney, and Inghram	85, 66	Dual electronic detector	Intensity measurement adaptable to fluctuating ion beams
Forrester and Whalley	75, 77	AC ion beam recording	Rapid panoramic scanning
Inghram, Hayden, and Hess	86	Vibrating reed electrometer	Intensity measurement

The first electrical recording of the ion beams was done by Thomson.[81] For this work he used a Faraday collector directly behind a defining slit and recorded the ion current with a Dolezalek type electrometer. Later Bleakney[35] replaced the electrostatic electrometer with an electrometer tube.

The use of multiple collectors was first reported by Straus.[66] They were necessary in his work, since he used the hot spark, and an integration of the ion beams was necessary. He measured the ion currents with Compton electrometers. A multiple collector system using direct coupled DC amplifiers for routine isotope measurement was reported by Nier, Ney, and Inghram.[85] The schematic diagram of their arrangement is shown in Fig. 18. With the dual collector shown at the left of this figure one ion beam is collected on the slit plate *S*, and the second which passes through this slit plate is collected on the collector *C*. These plates are surrounded by electron repelling fields of 22.5 volts to repel secondary electrons formed by the ion beam back to the collector. In alternative designs the secondary electron repellers are left out and secondary electron effects eliminated by placing a small magnet about the collector. Nier now uses a cup type collector at *C* to retain secondary positive ions formed by the primary beam. With the electronic amplifiers shown the ratio of the two ion currents is obtained by adjusting the attenuator *X* to the point where the balance meter *G* reads zero, when this is true the ratio of the two ion currents is given by the equation:

$$\frac{i_2}{i_1} = X\frac{R_1}{R_2}\left(\frac{G}{G+1}\right) \tag{40}$$

where X is the fraction of the total output voltage of the feed back amplifiers fed back to the electrometer circuit, R_1 and R_2 are the input resistors for currents i_1 and i_2, and G is the gain of the feed back amplifier. With such a circuit isotopic comparisons are easily made. The procedure is as follows. One sample is introduced and the attenuator *X* adjusted to balance. Then if a second sample is introduced which has exactly the same isotopic composition no change in balance will be observed. If there is a change the difference in the two samples is obtained by varying the attenuator to balance. The major advantage of this system over other electronic detector systems is that it operates independently of fluctuations in ion intensity.

The recording detector system was first reported by Smith, Lozier, Smith, and Bleakney.[83] Due to improvements in amplifier systems this advancement is now almost universally accepted in mass spectrometry.

The first use of an electron multiplier in a mass spectrometer was that of Cohen.[84] Unfortunately a description of the apparatus he used has never been published. In a multiplier used by Inghram and Rustad, the meter end of the multiplier was operated at ground potential so that the first plate of the multiplier, i.e., the plate corresponding to plate C in Fig. 18, was at −4000 volts. This potential is very convenient in the mass spectrometer application since it adds to the energy of the ion beam and hence gives a more efficient conversion of ion beam to electron beam. The difficulty with the electron multiplier system is that the vacuum in the multiplier is not constant, it changes with sample and time, so that the requirement of multiplier surfaces stable to many gases is very stringent. This type of detector will probably never be used for other than research purposes.

There is one variation of the electron multiplier application, which may become valuable in quantitative work. This system uses a fluorescent screen to collect the ion beam and a standard photomultiplier tube to record the fluorescence produced.

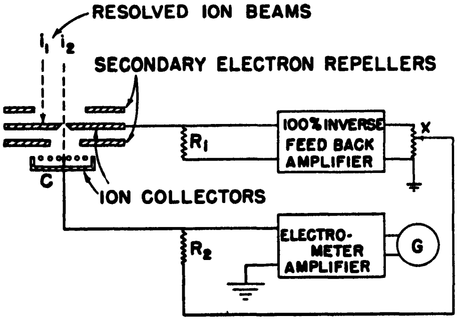


FIG. 18.—The dual detector system devised by Nier, Ney, and Inghram for comparison of isotopic compositions. The meter shown at G is a null detector. The ratio of the ion currents is obtained by adjustment of X to make G read zero.

A good deal of work has been done in adapting AC detecting systems,[75,77] but up to the present time no really quantitative work has been reported with pulsating ion beams.

The most recent advancement in detection systems has been the addition of the vibrating reed type of amplifier[86] developed by Palevsky, Swank, and Grenshik[88] in place of the DC amplifiers used previously. This system has a DC noise level ten times less than any of the previously used amplifiers and essentially no drift with time.

9. Electronic Components

It is not important here to discuss the various electronic circuits used in conjunction with mass spectroscopes. For example, there are as many different methods of controlling the magnetic fields as there are investigators in the field of mass spectroscopy. For details of electronic components the reader is referred to the references listed in Table II. It is sufficient to state that straightforward electronic components are available that meet most requirements.

IV. USES OF THE MASS SPECTROSCOPE

1. *Isotope Existence*

The first problem to which the mass spectrograph was applied was that of identifying the isotopic composition of the elements. The first element shown, by Thomson in 1913, to consist of more than one isotope was neon. Since that time 302 isotopes which occur by natural processes in terrestrial matter have been identified. This field is now practically complete. Only one naturally occurring isotope has been found in the last eight years.[89] Future work in the field of isotope existence must be done with spectrometers of exceedingly high sensitivity and low background. The existence of several isotopes which might be present in sufficient quantities to be detected by the mass spectrometer has been summarized in several semiempirical treatments.[90,91]

2. *Isotopic Abundances*

The second major application of the mass spectroscope was the problem of determining the isotopic composition of the elements. Both the mass spectrograph and the mass spectrometer are applicable to this problem, though the results obtained with the spectrometer are more accurate.

The procedure used in the case of the mass spectrograph is as follows: The element under investigation is run in the spectrograph to obtain its characteristic line spectrum (see spectrum (a) of Fig. 17). A microphotometer is then used to obtain the optical densities of the spectral lines. By comparing these optical densities with standard photometric density curves, the exposure and hence the isotopic composition of the element can be obtained.

There are a number of important errors inherent in such mass spectrographic measurement which are not encountered with mass spectrometers. The first arises from the fact that the trajectories of the ion beams of different mass are different. Thus, geometrical discriminations enter. Closely associated with this error is the fact that in all mass spectrographs the line widths of the spectral lines vary along the detector plate. The Mattauch machine has line widths which are proportional to the square root of the mass. The Dempster and Bainbridge machines have line widths which are functions of the distance from the position of optimum focus. Thus, in mass spectrographic measurement the widths must be taken into account in determining the isotopic abundance. A second error encountered with the mass spectrograph results from the fact that ions of different mass have different velocities. This introduces errors since ions of different velocity do not give the same photographic blackening. A third difficulty comes in due to the fact that the standard density curve varies according to whether the standards are printed with light, x-rays, or ions of different mass or velocity. The method most often used to obtain the standard density curve is the method of using line spectra obtained from elements whose isotopic composition is known from mass spectrometric measurements. This gives a standard density curve by means of which the isotopic composition of the unknown may be obtained. This, however, assumes that the previous mass spectrometric values are absolute. It is obvious from this discussion that since the most reliable spectrographic method uses a mass spectrometrically determined value to calibrate, that the latter is accepted as the most reliable. As mass spectrometrically determined values for isotopic abundances become available they will replace the earlier spectrographic values.

The mass spectrometric measurement has been developed by Dempster,[18,28] Bleakney,[35] Bainbridge,[50] and Nier[37] to its present high accuracy. With this method there are still a large number of uncertainties that limit the accuracy. Unless precautions are taken to eliminate as many as possible of these errors and to evaluate the others the results obtained are unreliable. The important sources of error are:

(1) Fractionation of the material in the process of introduction into the mass spectrometer.
(2) Fractionation in the ionization process.
(3) Space charge effects in the ion source.
(4) Fractionation in the ion source due to magnetic discriminations and voltage effects.
(5) Fractionation if the ion paths are not identical.
(6) Secondary effects associated with the collection of the ion currents.
(7) Nonlinearity or polarization of the ion current measuring circuit.
(8) Variation in the shape of the ion beams.

The first of these errors have been discussed in section III-2.

Possible fractionation in the ionization process results if the ionization efficiency for the two isotopes is different. The fact that such effects are detectable in hydrogen is illustrated by the results of Evans[92] showing that the ionization efficiencies of CH_4 are different from those of CH_3D. For this reason the ratio of the hydrogen isotopes in normal material has been determined by mixtures of separated isotopes.

The possible discrimination due to space charge has been suggested by Bainbridge[93] to explain discrepancies in the normal abundances of lithium. This possible effect can be minimized if the spectrometer is operated at such low electron and ion currents that space charge effects are negligible.

The systematic discriminations in the ion source due to magnetic effects has been dicussed in mathematical detail by Jordan and Coggeshall.[94,95] These discriminations can be eliminated by operating the ion source with all magnetic fields shielded out. In addition, errors due to voltage effects, i.e., the variation in the efficiency of the ion source as the potential across it is varied, can be eliminated by holding the accelerating voltage constant and bringing the ion beams of different mass to focus on the fixed collector by varying the analyzer magnetic field. These two conditions automatically fulfill the requirements of making ion trajectories identical throughout the machine. Most of the recently reported values have been obtained using this technique.

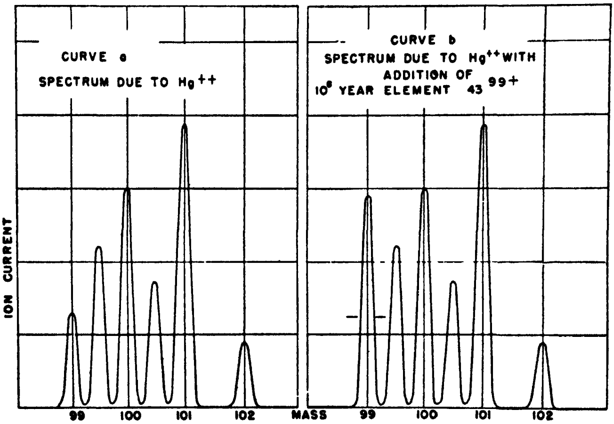


FIG. 19.—Spectra proving that the mass of the long lived fission isotope of technetium (element 43) is 99. Technetium is the lightest element in the atomic table which does not exist in nature.

One other important error encountered in mass spectrometer measurement is nonlinearity in the ion detector resistor, (R_1 and R_2 in Fig. 19). In general nonlinearities are proportional to the value of the resistance. Measurements on 10^{12} ohm resistors yield nonlinearities averaging about 20% in the range 1 millivolt to 10 volts. Obviously, under such conditions accurate isotopic abundances are impossible. The majority of resistors in the range of 10^{10} ohms are linear to better than 0.5%.

It is apparent from this discussion and section III-2 that the determination of normal isotopic abundances is by no means a simple job. It is one thing to measure the peak heights with a mass spectrometer. It is quite another matter to say that these values give an absolute measurement.

3. *Packing Fractions*

The third important task to which the mass spectrograph was put was that of determining the exact masses of the isotopes. It was thought at one time that the masses of all isotopes could be expressed as a multiple of a simple fundamental unit. The works of Aston, Bainbridge, Dempster, Bainbridge and Jordan, and Mattauch have proved that this is not exactly true, but that each isotope is characterized by a mass defect, i.e., difference from an integral mass. These mass defects are characterized by numbers called packing fractions which are defined by the equation:

$$PF = \frac{M - I}{I} \times 10^4 \tag{41}$$

where I is the nearest integer to M, and M is the exact mass of the isotope under investigation is units of one-sixteenth of the lighest oxygen isotope which is defined to be mass 16.0000.

It is of importance to note that the physical definition of mass is different from that used by the chemists. The physicist defines the atomic mass unit to be one-sixteenth that of the lightest isotope of oxygen, while the chemist defines it as one-sixteenth of the normal oxygen, which is about 1.000275 times larger than the physical unit. Thus, all atomic weights on the physical scale are larger than those on the chemical scale by this factor. The physical definition of mass is more accurate, since it is defined in terms of a definite mass, i.e., the mass of 0^{16}, while the chemical definition is ambigious in that it defines the mass unit in terms of "normal" oxygen. Dole[96] has shown that the abundance of the heavy isotopes in "normal" oxygen varies by 4%. This means that the chemical mass unit varies by 10 parts/million depending on the source of the "normal" oxygen.

Only double focus or velocity focusing spectrographs can be used for accurate packing fractions. It has been suggested that a direction focus machine might be used for this work.[97] It must be pointed out that the application of a single focus machine to packing fraction is very limited. Any determination to which the Frank-Condon[98,99] principle applies, i.e., where there is a splitting of a molecule, requires the more complicated double focusing instruments. A second factor which makes even the double focusing mass spectrometer inaccurate in the problem of packing fraction measurement is the fact that in the presence of an ion beam there are always surface polarization effects, which make the absolute magnitude of the potential applied to accelerate and deflect the ion beam uncertain.[100] These effects are appreciable even if pure gold surfaces are used. With the mass spectrograph, however, the packing fractions are determined by measuring distances along a photographic plate, and the values obtained are independent of any polarization effects in the electrostatic acceleration and analyzer system. For this reason packing fractions have been measured only on mass spectrographs.

There are four different methods which are used for determining packing fractions with the mass spectrograph. These are (1) the multiple charge doublet, (2) the molecular doublet, (3) the series shift bracket, and (4) the ratio bracket.

The multiple charge type of doublet is illustrated by Fig. 17 (b) obtained by Dempster. The left line of this doublet is produced by the triply charged titanium isotope of mass 48. The right hand line is produced by the singly charged oxygen isotope of mass 16. Referring to eq. (16) it is apparent that since the position of the line on the plate depends on the ratio of m/e, that if Ti^{48} had exactly three times the mass of 0^{16} its triply charged ion would fall directly on top of the 0^{16} line. The fact that it falls below, proves that its mass is less than three times that of 0^{16}. To determine the mass of Ti^{48} Dempster assumes that the mass difference of the isotopes of Ti are whole numbers. Thus, by measurement of the 47 to 48 and 48 to 49 distances the average dispersion at mass 48 is known. Using this dispersion and measuring the doublet distance, the mass of the Ti^{48} as compared to 0^{16}, which is defined as 16.00000 is immediately obtainable. As a typical example assume that the intervals 47–48 and 48–49 can be reproduced to 1 part in 200, and that the doublet Ti^{48}-0^{16} can be measured to 1 part in 20. Under these conditions the error in the dispersion owing to the assumption of integral mass units is negligible. However, measurement of the doublet distance to 1 part in 20 gives the mass of Ti^{48} to 1 part in 30,000. This example illustrates why the mass spectrograph can be used to give such accurate masses.

An example of molecular type of doublet is the doublet formed by the molecule $C^{12}H_4$ and the atom 0^{16}. Obviously, if isotopes had integral masses these two ions would have exactly the same mass and hence be indistinguishable in the mass spectrograph. The fact that they form a doublet which looks very similar to the Ti^{48}-0^{16} doublet just discussed proves that the masses are not integral and measurements of doublet separation again gives the mass defects. For this type of doublet the dispersion at mass 16 is obtained by taking the average of dispersion in the interval 15–16 and 16–17 using the best available masses for the ions of mass 15 and 17.

The series shift method is used whenever the masses to be compared form terms in a series whose mass difference is small compared to the mass of that series. An example of such a determination is the odd mass, even mass discrepancy in the isotopes of neodymium. The normal neodymium spectrum is shown in (a) of Fig. 17. The question as to whether the 143 mass is intermediate between 142 and 143 can be answered roughly from this spectrum by measurement of the intervals 142–143, 143–144, and 144–145. The difficulty with this determination is that these distances can be reproduced to, for example, only 1 part in 300. This relatively poor reproduction is due to field inhomogeneities. If, however, two exposures are taken with slightly different fields so that a second spectrum is formed displaced by approximately one mass unit, the accuracy of the measurement of the 142–143, 143–144, and 144–145 intervals is still 1 part in 300, but these intervals now represent a much smaller unit of mass than in the simple spectrum. For example, if the second set of lines are moved nine-tenths of a mass unit away from the first set, the accuracy of the measurement of the mass difference is increased by a factor of 10. This serves to illustrate the value of the series shift method.

The ratio shift method is used where the ratio of two masses to be determined can be compared with a known ratio. For example, Aston had determined the mass of F^{19} by this method. In this case doublets are formed by shifting the F^{19}-$C_2{}^{12}$ series so that it coincides approximately with the C^{12}-$CH_3{}^{15}$. Again the measurement involves measurement of a small fraction of a mass unit so that results are very accurate.

Even with the mass spectrograph there are a number of effects which must be carefully watched. Among these are such effects as inhomogeneities in the fields so that the results are a function of the position at which the beam strikes the spectrograph plate. Another is the fact that if the ion currents are too strong the photographic plate, which is an insulator, takes on a charge. In the case of close doublets the effect can increase the apparent distance between the doublet lines. In making very heavy deposits for radioisotope mass assignment such effects are very detectable. Again it is apparent that only by careful evaluation of the type of result that is desired can proper corrections and conditions be obtained.

4. Determination of the Mass of Radioactive Isotopes

The masses of some of the artificially produced radioactive isotopes have been determined by the mass spectroscope.[101–109] This method is usually used when other nuclear reaction data fail. For example the rare earth element europium has two naturally occurring isotopes at masses 151 and 153. By slow neutron bombardment 9.2-hour and 6-year activities are induced. The reactions are simple $(n\gamma)$ reactions.

From the reaction it is impossible to say whether these activities are due to active isotopes of mass 152 or 154 or both.

There are three different methods of applying the mass spectroscope to this problem. They are (1) the separation of active materials into their various mass components with a mass spectrograph so that the activity can be located among the separated isotopes, (2) the separation of stable isotopes which are subsequently activated, and (3) study of the normal spectra of activated isotopes to locate abnormal lines. Of these methods only the first and third are applicable to fission product isotopes. The limits of this type of application can best be illustrated by an explanation of the first of these methods.

A sample containing the activity in question and an inert material to act as mass standard is run in a mass spectrograph to obtain a photographic plate on which the separated isotopes are deposited. Activities with half lives of less than 10 hours are then located in the following way. The plate with active deposit is placed on one side of a heavy slit and a Geiger counter on the other. By moving the plate across the slit maxima in activity are observed at the positions of the active lines. The mass is then determined by development of the original plate to show the mass standards.

This technique is best for short half lives, i.e., from 15 minutes to 10 hours. To obtain definite results the active deposit must have a decay rate of greater than 100 disintegrations/minute. This gives a counter rate of $\sim$ 10 disintegrations/minute, depending on decay particles and energies.

The second and more spectacular method is to place the plate containing the activity face to face with a second photographic plate. The radioactive particles emitted by disintegration of the active deposit on the first plate will give rise to a developable image on the second plate. Development of both plates will then show the mass standards and the active lines on the original plate and only the active lines on the second or radio autograph plate. An example of this method is shown in Fig. 17, spectra (c and d). In this case the original spectrum shows the normal lines of neodymium plus two lines due to element 61 at masses 147 and 149. That these lines are active is proved by the radio autograph shown in spectrum d. To obtain a good developable image on the autograph plate about $10^4\beta$ disintegrations/sq. mm. of active deposit are required. This means that a deposit of 10^4 active atoms/sq. mm. of a short lived material is sufficient for a mass assignment. This method is the most practical method for half lives of from 10 hours to 50 years.

The third method completely neglects the fact that the isotopes are radioactive. The method is simply to run the material as one would to determine the normal isotopic structure, and to determine the mass from the primary spectrum obtained. This method requires more material than either of the above methods, but is actually more efficient for half lives of greater than 50 years than either of the above mentioned methods. An example of this method is the determination of the mass of long lived fission technetium[110] (element 43). This is one of the elements that does not exist in nature. The method is illustrated by Fig. 19. Curve (a) shows the normal mass spectra in the mass range 98–103 with the peaks of doubly charged mercury added to serve as mass standard. Curve (b) is the same region plotted after the long lived fission technetium was added to the spectrum. The conclusion that the mass of long lived technetium is 99 is immediately apparent.

These methods of mass assignment have been used for over thirty active isotopes whose masses were previously unknown. There is still a large amount of work to be done in this field.

5. Neutron Absorption Cross Sections

Another recent application of the mass spectroscope has been the determination of the neutron absorption cross sections of stable isotopes.[111–113] The application[113] is illustrated by the curves shown in Fig. 20. The curve on the left shows the mass spectrum observed for normal mercury and the curve on the right shows the spectrum for mercury which has been submitted to neutron bombardment in a chain reacting pile. Three marked changes are immediately apparent. One is that 23% of the mercury isotope of mass 196 has disappeared, the second is that 19% of the isotope at mass 199 has disappeared, and the third is that the isotope at mass 200 has increased by 14%. The explanation of the change is as follows. The neutrons (mass 1) bombarding the mercury sample are absorbed in the mercury isotopes of mass 196 and 199, changing them to isotopes of mercury with masses 197 and 200. Computation will show that the decrease in the peak height at mass 199 is just equal to the increase at mass 200, explaining these two changes. The mercury isotope at mass 197 is not present because it is radioactive and decays to gold. From this curve it is thus apparent that the big neutron absorbers in mercury are the isotopes of mass 196 and 199.

6. Gas Analysis

One of the most important commercial applications of the mass spectrometer is the analysis of gas mixtures.[4–9] For example, the analysis of a mixture of isobutane C_4H_{10} and normal butane C_4H_{10} is a simple matter with the mass spectrometer, while the same analysis by organic chemical methods is tedious.

The spectra characteristics of normal and isobutane are shown in Fig. 21 along with the chemical structure for each. It is immediately apparent that the ratio of the 57 to 58 peaks is quite different for the two compounds. The small peak at mass 59 is due to the presence of the C^{13} isotope. Obviously, if the ratio of the 57 to 58 peaks is between these two cases for an unknown butane sample the abundances of the two butanes can be computed by the solution of two simultaneous equations. In general, for an n component mixture, n simultaneous equations must be solved. This is not difficult, since electronic apparatus has been worked out for the rapid analysis of such systems.[114]

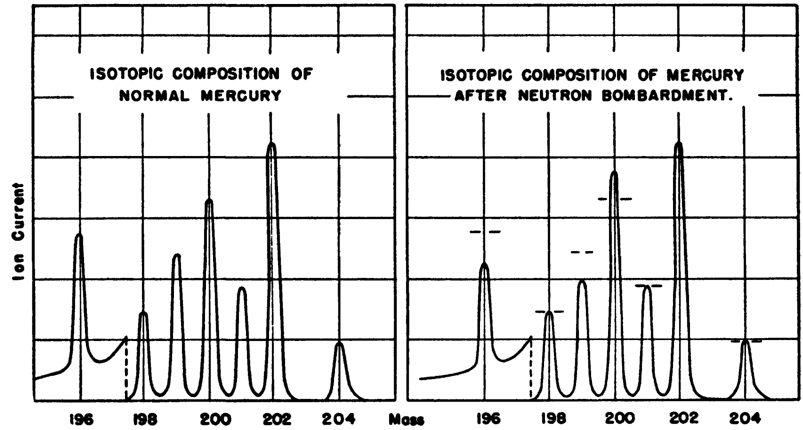


FIG. 20.—Mass spectrometer traces of the isotopes of mercury before and after bombardment in a chain reacting pile. The decrease in the peaks at mass 196 and 199 in the bombarded sample is due to the neutron absorption of these isotopes.

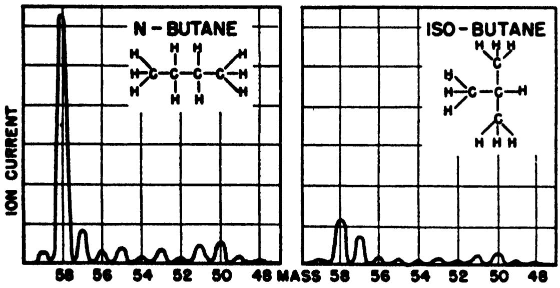


FIG. 21.—Mass spectra characteristic of normal and isobutanes.

There are a number of analyses, however, where the calculation is much simpler. For example in a mixture of normal butane and pentane the series of peaks at mass 72 is due to pentane alone while those at 58 are due to both pentane and *n*-butane. In this case, the method of solution is to subtract the pentane spectra from the butane peaks. The remaining peaks at mass 58 will be due to butane alone. Thus, the problem reduces to a simple subtraction.

TABLE IV. Typical analysis of a five component hydrocarbon mixture by the mass spectrometer.[5]

	Manometer %	Mass spectrometer %	Difference
Methane	19.1	19.5	0.4
Ethane	37 7	38.0	0.3
Propane.	39 6	39.0	−0.6
Isobutane	2.1	2.0	−0.1
n-Butane	1.5	1.5	0.0
Methane	39 2	40.1	0.9
Ethane.	52.3	51.7	−0.6
Propane	3.7	3.2	−0.5
Isobutane	3.1	3.2	0.1
n-Butane	1 7	1 6	−0 1
Methane	42 7	43 5	0 8
Ethane	0	0 4	0 4
Propane	52 6	51 3	−1.3
Isobutane	3 0	3 0	0 0
n-butane	1 7	1 8	0 1

As an example of the use of the mass spectrometer in the analysis of a five component mixture Table IV shows the results of Hoover and Washburn for three such mixtures.[7] The accuracy of analysis according to this table is of the order of 1%.

One of the remarkable advantages of this type of analysis is the time required for an analysis. Young[115] has tabulated the average time required for routine analysis of 1580 samples. He reports that it takes an average of 1.31 man hours/sample for the analysis of three to four component mixtures and 1.90 man hours/sample for mixtures containing ten to twenty components each.

There are several machines available commercially designed specifically for the analysis of gas mixtures. The companies supplying these machines are listed in section V of this article.

7. *Solid Analysis*

The mass spectrograph has recently been applied by Dempster[116] to chemical analysis of solid samples, in much the same way that has been used since 1910. In chemical analysis it has two marked advantages over the optical spectrograph. (1) The mass spectrum characteristic of each element is much simpler than the optical spectrum. For example, the first order mass spectrum of iron contains only four lines whereas the first order optical spectrum contains 959 "principal" lines, (2) the mass spectrum has no blind spots as in the case of the optical spectrograph. Any element which is present is recorded. The major disadvantage of the instrument is that it involves vacuum techniques; however, with the advancements in vacuum techniques this is not a serious drawback. The electron microscope is a good example of a vacuum instrument which is now used routinely in many laboratories.

Undoubtedly the most universal source available at the present time for this work is the vacuum spark source. Typical results obtained by Dempster with this source are given in Table V, which gives the minimum amount of impurities detected when a 1 mg. sample of uranium was consumed by the vacuum spark. The time required for such an exposure is about 20 seconds.

TABLE V. Impurities detectable from a 1 mg. sample of uranium.

Impurity	Be	B	C	N	O	F	Va	Mg
Impurity in parts/million	0 04	0 12	0 8	0 24	0 08	0.09	0.05	0.35

The amount of impurity in a sample is determined exactly as with the optical spectrograph, i.e., by photometric measurement of the line density. Also as in optical spectrography calibration is accomplished by running known standards.

In general, the sensitivity of the mass spectrograph decreases as the mass increases. This is due to the fact that the photographic density of an image decreases as the velocity of the ion decreases. However, by increasing the exposure time these should be detected equally well.

In the future this application of the apparatus may develop rapidly, especially if machines become available. The time between the development of optical spectroscopy and the application to chemical analysis was fifty years. Probably within that length of time the use of mass spectrographs for solid analysis will be quite commonplace.

8. Leak Detection

The mass spectrometer is the most sensitive leak detector known.[11,117] The machines now available commercially are able to detect the presence of one leak in the presence of 200,000 other equal leaks.

The method is illustrated schematically in Fig. 22. As is seen from this diagram, the inlet to the mass spectrometer is connected to the system under test so that the gas in the spectrometer is characteristic of that system. To detect leaks, helium is sprayed externally onto the test system. Whenever the helium gas comes in contact with a leak it passes through that leak and into the detecting mass spectrometer. If the mass spectrometer is set to record the mass 4 ion current, i.e., helium, a current will be detected and the presence of a leak indicated.

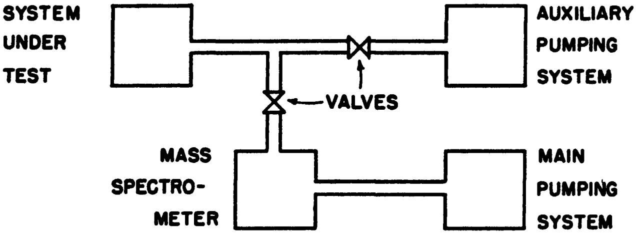


FIG. 22.—Schematic diagram illustrating the application of the mass spectrometer to leak detection. As a leak detector, the mass spectrometer can pick out one leak in the presence of 200,000 other equal leaks.

The intensity of the current recorded in the mass spectrometer is a measure of the size of the leak.

In a small system the auxiliary pumping system is not necessary; the pumping system of the spectrometer itself is sufficient for evacuation. However, when large systems, or systems with large leaks are encountered, auxiliary pumping equipment is necessary to reduce the pressure in the system under test to usable values, and to increase the speed of response.

The use of helium is indicated since there is very little helium in the atmosphere (about 1 part in 500,000). The method of leak testing however, is not limited to helium because hydrogen, xenon, argon, carbon dioxide, volatile hydrocarbons, etc., can be used when necessary.

There are at the present time three commercial models of leak detectors available for laboratory and industrial purposes. All these machines are constructed for use of helium probes only.

9. Other Applications of the Mass Spectrometer

The above applications are illustrative of the type of problems handled by the mass spectroscope. There is not enough time or need in an article such as this to give discussions of each of the many applications which have been made. Hence, having illustrated the techniques that are used it will be sufficient to list a number of other problems in which the spectroscope has provided valuable information. The list below includes a number of these applications along with references to recent results obtained in these fields.

(1) Biological and chemical tracers problems[10,118]
(2) Free radical phenomena[119,120]
(3) Age of the earth[121]
(4) Age of the elements[122,123]
(5) Metastable ions[124]
(6) Variation of isotopic composition in nature[125]
(7) Gases given off in outgassing processes
(8) Ionization phenomena and bond strengths[9,126]
(9) Isotope equilibrium and separation techniques[127]
(10) Kinetic theory application[128,129]
(11) Radioactive half lives[130–132]

V. COMMERCIALLY AVAILABLE MASS SPECTROMETERS

There are four commercial companies which are now supplying special mass spectrometers for industrial and research purposes. These companies are:

(1) The Consolidated Engineering Corp.
620 North Lake Avenue
Pasadena 4, Calif.
(2) The General Electric Co.
Schenectady, N. Y.
(3) Vacuum-Electronic Engineering Co.
316 37th Street
Brooklyn, N. Y.
(4) Process & Instruments
60 Greenpoint Avenue
Brooklyn, N. Y.

Unfortunately there is at the present time no commercially available mass spectrograph on the market.

REFERENCES

1. Thomson, J. J. Rays of Positive Electricity and Their Application to Chemical Analyses. Longmans Green, New York, 1913, p. 20.
2. Aston, F. W. *Phil. Mag.*, **45**, 941 (1923).

3. Pollard, E., and Davison, W. L. Applied Nuclear Physics. Wiley, New York, 1942, p. 191.
4. Tate, J. T., Smith, D. T., and Vaughan, A. L. *Phys. Rev.*, **48,** 525 (1935).
5. Hoover, H. H., and Washburn, H. W. *Proc. Calif. Natural Gas Assoc.* 16th Annual Fall Meeting (1941).
6. Hipple, J. A. *J. Appl. Phys.*, **13,** 551 (1942).
7. Washburn, H. W., Wiley, H. F., and Berry, C. E. *Industr. Engng. Chem.* (*Analyt. Edit.*), **17,** 74 (1945).
8. Washburn, H. W. Physical Methods in Chemical and Metallurgical Analyses. Edited by W. Berl, Academic Press, New York, in press.
9. Mariner, T., and Bleakney, W. *Phys. Rev.*, **72,** 792 (1947).
10. Wilson, D. W., Nier, A. O., and Reiman, S. P. Preparation and Measurement of Isotopic Tracers. J. W. Edwards, Ann Arbor, Mich., 1946.
11. Nier, A. O., Stevens, C. M., Hustrulid, A., and Abbot, T. A. *J. Appl. Phys.*, **18,** 30 (1947).
12. Nier, A. O. *Bibliog. Sci. Ind. Repts.*, PB33031 U. S. Dept. of Commerce (1946).
13. Coggeshall, N. D. *J. Appl. Phys.*, **18,** 855 (1947).
14. Rogers, F. T. *Rev. Sci. Instrum.*, **11,** 19 (1940).
15. Herzog, R. *Z. Phys.*, **89,** 447 (1934).
16. Bainbridge, K. T., and Jordan, E. B. *Phys. Rev.*, **50,** 282 (1936).
17. Nier, A. O. *Rev. Sci. Instrum.*, **11,** 212 (1940).
18. Dempster, A. J. *Phys. Rev.*, **11,** 316 (1918).
19. Mattauch, J., and Herzog, R. *Z. Phys.*, 89, 786 (1934).
20. Bartky, W., and Dempster, A. J. *Phys. Rev.*, **33,** 1019 (1929).
21. Dempster, A. J. *Phys. Rev.*, **51,** 67 (1937).
22. Cartan, L. *J. phys. radium*, **8,** 453 (1934).
23. Hutter, R. G. E. *Phys. Rev.*, **12,** 19 (1944).
24. Stephens, W. E. *Phys. Rev.*, **45,** 513 (1934).
25. Thomson, J. J. Rays of Positive Electricity and Their Application to Chemical Analysis. 1913.
26. Aston, F. W. *Phil. Mag.*, **47,** 385 (1923).
27. Moon, P. B., and Oliphant, M. L. *Proc. Roy. Soc.*, **A137,** 463 (1932).
28. Dempster, A. J. *Phys. Rev.*, **20,** 631 (1922).
29. Dempster, A. J. *Proc. Amer. Phil. Soc.*, **75,** 755 (1935).
30. Wall, R. F. *National Electronic Conference* Chicago, Nov. 1947.
31. Smith, L. P. *Phys. Rev.*, **72,** 153 (F5) (1947).
32. Koch, J., and Bendt-Nielsen, B. *Kgl. Danske Videnskab. Selskab, Mat.-fys. Medd.*, **21,** (No. 8) (1944).
33. Langmuir, I., and Kingdon, K. H. *Proc. Roy. Soc.*, **A107,** 61 (1925).
34. Shaw, A. E. *Bibliog. Sci. Ind. Repts.*, PB52763 U. S. Dept. of Commerce (1946).
35. Bleakney, W. *Phys. Rev.*, **40,** 496 (1932).
36. Tate, J. T., and Smith, P. T. *Phys. Rev.*, **46,** 773 (1934).
37. Nier, A. O. *Rev. Sci. Instrum.*, **18,** 398 (1947).
38. Penning, F. M. *Physica*, **3,** 873 (1936); **4,** 71 (1937).
39. Knudsen, M. *Ann. Phys. Lpz.*, IV, **28,** 75–131 (1909).
40. Knudsen, M. *Ann. Phys. Lpz.*, IV, **28,** 999–1016 (1909).
41. Snoluchowski, M. *Ann. Phys. Lpz* , **33,** 1559 (1910).
42. Kennard, E. H. Kinetic Theory of Gases. McGraw Hill, New York, pp. 294.
43. Honig, R. E. *J. Appl. Phys.*, **16,** 646–654 (1945).

44. Nier, A. O., Ney, E. P., and Inghram, M. G. *Rev. Sci. Instrum.*, **18,** 191 (1947).
45. Mulliken, J. *J. Amer. Chem. Soc.*, **44,** 2387 (1922); **45,** 1592 (1923).
46. Aston, F. W. *Phil. Mag.*, **38,** 709 (1919).
47. Costa, J. L. *Ann. Phys., Paris*, **4,** 425 (1925).
48. Aston, F. W. *Proc. Roy. Soc.*, **A115,** 487 (1927).
49. Bleakney, W. *Phys. Rev.*, **34,** 157 (1929); 35, 139 (1930).
50. Bainbridge, K. T. *Phys. Rev.*, **40,** 130 (1932).
51. Smythe, W. R. *Phys. Rev.*, **28,** 1275 (1926).
52. Mattauch, J. *Phys. Z.*, **33,** 899 (1932).
53. Hintersberger, H., and Mattauch, J. *Z. Phys.*, **106,** 279 (1937).
54. Oliphant, M. L., Shire, E. S., and Crowther, B. M. *Proc. Roy. Soc.*, **A146,** 922 (1934).
55. Bondy, H., Johannsen, G., and Popper, K. *Z. Phys.*, **95,** 46 (1935).
56. Smythe, W. R., Rumbaugh, L. H., and West, S. S. *Phys. Rev.*, **45,** 724 (1934).
57. Smythe, W. R., and Hemmendinger, A. *Phys. Rev.*, **51,** 178 (1937).
58. Rumbaugh, L. H. *Phys. Rev.*, **49,** 882 (1936).
59. Sampson, M. B., and Bleakney, W. *Phys. Rev.*, **50,** 456 (1936).
60. Aston, F. W. *Proc. Roy. Soc.*, **163,** 391 (1937).
61. Mattauch, J. *Phys. Rev.*, **50,** 617 (1936).
62. Nier, A. O. *Phys. Rev.*, **50,** 1041 (1936).
63. Bleakney, W., and Hipple, J. A. *Phys. Rev.*, **53,** 521 (1938).
64. Smythe, H. D. *Phys. Rev.*, **25,** 452 (1925).
65. Jordan, E. B. *Phys. Rev.*, **57,** 1072 (1940).
66. Strauss, H. A. *Phys. Rev.*, **59,** 430 (1941).
67. Brown, H., Mitchel, J., and Fowler, R. D. *Rev. Sci. Instrum.*, **12,** 435 (1941).
68. Hipple, J., Grove, D., and Hickam, W. *Rev. Sci. Instrum.*, **16,** 69 (1945).
69. Coggeshall, N. D., and Jordan, E. B. *Rev. Sci. Instrum.*, **14,** 125 (1943).
70. Nier, A. O., Inghram, M. G., and Stevens, C. *Bibliog. Sci. Ind. Repts.*, PB52797 U. S. Dept. of Commerce (1943).
71. Taylor, J. E. *Rev. Sci. Instrum.*, **15,** 1 (1944).
72. Thomas, H. A., Williams, T. W., and Hipple, J. A. *Rev. Sci. Instrum.*, **17, 368** (1946).
73. Thode, H. G., Graham, R. L., and Ziegler, J. A. *Canad. J. Res.*, **B23,** 40 (1945).
74. Stephens, W. E. *Phys. Rev.*, **69,** 691 (1946).
75. Forester, A. T., and Whalley, W. B. *Rev. Sci. Instrum.*, **17,** 549 (1946).
76. Shaw, A. E., and Rall, W. *Rev. Sci. Instrum*, **18,** 278 (1947).
77. Siri, W. *Rev. Sci. Instrum.*, **18,** 540 (1947).
78. Thomson, J. J. *Phil. Mag.*, **13,** 561 (1907).
79. Dechend, H. V., and Hammer, W. *Proc. Heidelberg Acad. Soc.*, **21,** 12 (1910).
80. Thomson, J. J. *Phil. Mag.*, **21,** 225 (1911).
81. Thomson, J. J. Rays of Positive Electricity. 2nd ed., 1921, p. 120.
82. Metcalf, G. F., and Thomson, B. J. *Phys. Rev.*, **36,** 1489 (1930).
83. Smith, P. T., Lozier, W. W., Smith, L. G. and Bleakney, W. *Rev. Sci. Instrum.*, **8,** 51 (1937).
84. Cohen, A. *Phys. Rev.*, **63,** 219 (1943).
85. Nier, A. O., Ney, E. P., and Inghram, M. G. *Rev. Sci. Instrum.*, **18,** 294 (1947).
86. Inghram, M. G., Hayden, R. J., and Hess, D. C. *Phys. Rev.*, **72,** 349 (1947).
87. Aston, F. W. Mass Spectra and Isotopes. 2nd ed., Edward Arnold, London, 1942, p. 89.
88. Palevsky, H., Swank, R. K., and Grenchik, R. *Rev. Sci. Instrum.*, **18,** 298 (1947).

89. Inghram, M. G., Hess, D. C., and Hayden, R. J. *Phys. Rev.*, **72**, 967 (1947)
90. Feenberg, E. *Rev. Mod. Phys.*, **19**, 239 (1947).
91. Kohlman, T. P. *Phys. Rev.*, **73**, 16 (1948).
92. Evans, M. W., Bauer, N., and Beach, J. Y. *J. Chem. Phys.*, **14**, 701 (1946).
93. Bainbridge, K. T. *J. Franklin Inst.*, **212**, 317 (1931).
94. Jordan, E. B., Coggeshall, N. D. *J. Appl. Phys.*, **13**, 539 (1942).
95. Coggeshall, N. D. *J. Chem. Phys.*, **12**, 19 (1944).
96. Dole, M., and Slobrod, R. L. *J. Amer. Chem. Soc.*, **62**, 471 (1940).
97. Ney, E. P., and Mann, A. K. *Phys. Rev.*, **71**, 835 (1947).
98. Frank, J. *Trans. Faraday Soc.*, **21**, 536 (1926).
99. Condon, E. U. *Phys. Rev.*, **28**, 1182 (1926); **32**, 858 (1928).
100. Aston, F. W. Mass Spectra and Isotopes. 2nd ed., Edward Arnold, London, 1942, p. 75.
101. Inghram, M. G., and Hayden, R. J. *Phys. Rev.*, **71**, 130 (1947).
102. Inghram, M. G., Hayden, R. J., and Hess, D. C. *Phys. Rev.*, **71**, 27 (1947).
103. Inghram, M. G., Hayden, R. J., and Hess, D. C. *Phys. Rev.*, **71**, 643 (1947).
104. Inghram, M. G., Hess, D. C., Hayden, R. J., and Parker, G. W. *Phys. Rev.*, **71**, 743 (1947).
105. Parker, G. W., Lantz, P. M., and Inghram, M. G. *Phys. Rev.*, **72**, 85 (1947).
106. Hayden, R. J. *Phys. Rev.*, **74** (1948).
107. Helmholtz, A. C. *Phys. Rev.*, **70**, 982 (1946).
108. Nier, A. O., Booth, E. T., Dunning, J. R., and Grosse, A. V. *Phys. Rev.*, **57**, 748 (1940).
109. Inghram, M. G., Shaw, A. E., and Hess, D. C. *Phys. Rev.*, **72**, 515 (1947).
110. Inghram, M. G., Hayden, R. J., and Hess, D. C. *Phys. Rev.*, **72**, 1269 (1947).
111. Lapp, R. E., Van Horn, J. R., and Dempster, A. J. *Phys. Rev.*, **71**, 745 (1947).
112. Dempster, A. J. *Phys. Rev.*, **71**, 829 (1947).
113. Inghram, M. G., Hess, D. C., and Hayden, R. J. *Phys. Rev.*, **71**, 561 (1947).
114. Berry, C. E. *Phys. Rev.*, **69**, 135 (A) (1946).
115. Young, W. S. *Nat. Petroleum News*, **38**, R 212 (1946).
116. Dempster, A. J. *Bibliog. Sci. Ind. Rep.*, MDDC **370**, U. S. Dept. of Commerce (1946).
117. Jacobs, R., and Zuhr, R. *Rev. Sci. Instrum.*, **18**, 36 (1947).
118. Rittenberg, D. *J. Appl. Phys.*, **13**, 561 (1942).
119. Leiger, E., and Urey, H. C. *J. Amer. Chem. Soc.*, **64**, 994 (1942).
120. Elterton, G. C. *J. Chem. Phys.*, **10**, 403 (1942).
121. Nier, A. O. *Phys. Rev.*, **55**, 150 (1939).
122. Brown, H. *Phys. Rev.*, **72**, 347 (1947).
123. Brown, H., and Inghram, M. G. *Phys. Rev.*, **72**, 349 (1947).
124. Hipple, J. A. *Phys. Rev.*, **71**, 594 (1947).
125. Murphy, B. E., and Nier, A. O. *Phys. Rev.*, **59**, 771 (1941).
126. Hagstrum, H. D. *Phys. Rev.*, **72**, 947 (1947).
127. Thode, H. G., Graham, R. L., and Ziegler, J. A. *Canad. J. Res.*, **B23**, 40 (1945).
128. Ney, E. P., and Armstead, F. C. *Phys. Rev.*, **71**, 14 (1947).
129. Murphy, B. F. *Phys. Rev.*, **72**, 834 (1947).
130. Chamberlin, O., Williams, D., and Yuster, P. *Phys. Rev.*, **70**, 580 (1946).
131. Norris, L. D., and Inghram, M. G. *Phys. Rev.*, **70**, 772 (1946).
132. McMillan, E. M. *Phys. Rev.*, **72**, 591 (1947).

Particle Accelerators

M. STANLEY LIVINGSTON*

Brookhaven National Laboratory,† *Upton, L. I., New York*

CONTENTS

		Page
I.	Introduction	269
II.	Direct Voltage Generators	271
III.	Resonance Accelerators: The Cyclotron	271
IV.	Induction Accelerators: The Betatron	278
V.	Principles of Acceleration to High Energies	281
	1. Relativistic Equations of Motion	282
	2. Principles of Phase Stability	289
VI.	The Synchrotron	294
VII.	The Synchro-cyclotron	300
VIII.	The Linear Accelerator	306
IX	Future Possibilities: The Proton Synchrotron	312
	References	315

I. INTRODUCTION

The usefulness of high voltage particle accelerators in nuclear research is unquestioned. From the prewar fund of knowledge of nuclear physics came the necessary scientific facts and theories which made possible the unusually rapid expansion in this field, and which culminated in the development of the atomic bomb. To this fund of knowledge particle accelerators have contributed no small amount. It is clear that future developments in the application of atomic energy to other, peacetime problems will require still more fundamental knowledge of the atomic nucleus. It has been stated frequently by informed scientists that present developments have largely used up the accumulated fund of knowledge. An intensive program of research is now needed to stimulate and support further applications.

One of the most significant gaps in our present knowledge is the exact nature of nuclear forces, the forces between protons and neutrons which bind nuclei and which store atomic energy. Theoretical physicists now believe that these forces are associated with the creation and absorption of mesons, those particles of mass intermediate between proton and electron, which have been observed until recently only as secondary products of high energy cosmic rays. It seems essential, therefore, to produce mesons in the laboratory, so that their properties and interactions with nuclei can be studied under controlled conditions. This will require particle energies of many hundred million electron volts. So, a new field of nuclear research is being mapped out, involving the development of machines capable of producing charged particles, both positive and negative, having energies hundreds of times greater than those available from prewar accelerators.

---

* On leave from Massachusetts Institute of Technology.

† Work done in part under the auspices of the Atomic Energy Commission.

Particle accelerators are applications of the most fundamental branch of electronics, that which deals with the motions of ions and electrons in magnetic and electric fields. The development of accelerators has paralleled and sometimes paced progress in the electronics industry. The first ion accelerators were simple applications of high direct voltage to evacuated discharge tubes; they stimulated the development of high voltage x-ray machines. Focusing requirements forced intensive studies and improvements in the field of ion and electron optics. Next came the magnetic accelerators which employed high frequency electric fields to produce multiple accelerations and which had the many advantages associated with circular orbits. Certainly this experience was of significance in the radar and magnetron developments during the past war. Now the accumulated experience in the electronics of high frequencies and pulsed circuits is feeding back new techniques and concepts into the accelerator design field. Modern particle accelerators are utilizing much of this experience and are progressing still farther in adapting the more sophisticated implications in the equations of motion. The most significant development is a new class of accelerators based upon the principle of "phase stability," which makes it possible to accelerate particles synchronously through hundreds of thousands of small accelerations in such a manner that the orbits are stable and intensity is preserved. With these accelerators it should ultimately be possible to obtain energies in the billion electron volt range. Such machines will be large and expensive, and will require far more engineering design and development than is available in most academic research laboratories. In this development the engineering profession must make major contributions. The variety and scope of the engineering problems involved will become evident in the discussion to follow.

The purpose of this paper is to describe the physical principles involved in several of the more important types of electronuclear machines and in particular the new phase stable accelerators. It will attempt to show the advantages and limitations of the several machines, to estimate the energy limits and extrapolate to some possible future developments.

II. DIRECT VOLTAGE GENERATORS

The first instrument for acceleration of ions which produced nuclear disintegrations[1] used a voltage multiplier circuit of condensers and rectifiers to obtain a high potential which was applied to an evacuated discharge tube. Ions were accelerated through a series of electrodes by potentials of up to 700 kilovolts. Later developments of the voltage multiplier circuit by the Phillips Lamp Works at Eindhoven have extended this technique to over 1 million volts. Several other methods of producing direct voltages have also been used. The most successful and practical device is the electrostatic generator originated by Van de Graaff.[2] The modern electrostatic generator[3,4] consists of a large spherical high voltage terminal supported by insulators inside a pressure housing. Charge is sprayed on a moving belt which carries the charge to the insulated terminal and raises its potential; ions or electrons are accelerated through a long vacuum tube extending from the terminal to the grounded base. Insulation breakdown and corona discharge have limited this technique at present to around four million volts, but developments in progress show promise of somewhat higher potentials. It is not within the scope of this article to attempt a detailed description of such direct voltage generators. They have been and still are important tools for research, particularly where precise control of energy is required. In this discussion, however, they are cited only to provide a point of contrast for the resonance type accelerators which seem destined to be the supervoltage machines of the future.

III. RESONANCE ACCELERATORS: THE CYCLOTRON

Introduction of the concept of multiple acceleration of particles by an oscillating electric field has made it possible to obtain high energies without the necessity of high potentials and the consequent limitations of insulation breakdown. The first use of this principle was by Wideröe[5] in 1928. His apparatus was an elementary linear accelerator with two tubular electrodes in line on which was impressed an oscillating electric field. With this arrangement he was able to accelerate sodium and potassium ions to twice the energy available in a single traversal of the electric field. The next step was the extension of the concept of multiple acceleration to particles moving in circular orbits in a magnetic field, the "magnetic resonance accelerator" now known as the "cyclotron." This principle was first proposed by Lawrence[6] in 1930.

The cyclotron has been the most successful resonance accelerator, and the name has become so well known that it is often loosely applied to designate any type of particle accelerator. In the cyclotron a uniform magnetic field is used to constrain the ions in circular orbits so as to cause them to pass many times through the same set of electrodes in resonance with an oscillating electric field. The first practical machine was completed at the University of California in 1932[7] and was the first instrument to accelerate charged particles to 1 Mev (million electron volts) energy. Continued development to larger sizes and higher energies has culminated in the modern cyclotrons of 10–20 Mev energy which are described most completely in a summary article published in 1944.[8] Since this last reference includes a complete bibliography, and a detailed discussion of the techniques, only a brief survey of the physical principles and limitations will be presented here.

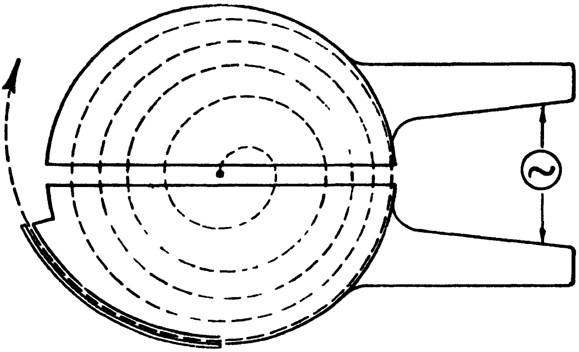


FIG. 1.—Schematic representation of the cyclotron. Ions produced at the center of a vacuum chamber are accelerated by high frequency electric fields between two semicircular, hollow electrodes and bent into circular paths by a uniform magnetic field normal to the plane of the ion orbits When they reach the periphery they are deflected outward by an auxiliary electric field.

The cyclotron accelerates positive ions many times through the same small radio frequency field to energies which are hundreds of times greater than the maximum voltages applied to the electrodes. (See Fig. 1.) Ions are produced between two semicircular hollow electrodes (called "D's" because of their shape) in the center of a vacuum chamber placed between the poles of a large electromagnet. The ions are accelerated by the radio frequency field between electrodes into the field-free region inside one of these electrodes and are deflected in a circular path by the uniform magnetic field so that they return to the diametral gap between the electrodes at some later time. For the condition of resonance the magnetic field and frequency are adjusted so that the time required for an ion to complete a semicircular path in the magnetic field is equal to the time for reversal of the radio frequency field. Under these conditions the ions find an accelerating field and obtain additional energy each time they cross the gap between electrodes. They travel in wider and wider semicircular paths until they reach the periphery of the electrodes, at which point they are deflected outward against a target.

The most useful particles for nuclear disintegrations have been ions of light and heavy hydrogen (protons and deuterons) and helium. An ion of mass m, charge e, and velocity v moving perpendicular to a magnetic field B will experience a force Bev, which is normal to the direction of the field and to the direction of motion. This produces motion in a circular path of radius r, such that:

$$Bev = \frac{mv^2}{r} \qquad \text{or} \qquad \omega = \frac{v}{r} = \frac{eB}{m} \tag{1}$$

The angular velocity of motion of the ions, ω, is constant as long as e, m and B are constant. This is the well-known relation for circular motion in a uniform magnetic field. The frequency of rotation of the ion, $f_i = \frac{\omega}{2\pi}$, is constant and independent of the linear velocity or radius of path. In the cyclotron the frequency of the oscillating electric field is set equal to this ion rotation frequency:

$$f = \frac{1}{2\pi}\frac{e}{m}B \tag{2}$$

This can be evaluated for the e/m values characteristic of light ions as:

Protons:	f (Mc) $= 1.52B$ (kilogauss)
Deuterons:	f (Mc) $= 0.76B$ (kilogauss)
Alphas (He^{++}):	f (Mc) $= 0.76B$ (kilogauss)

An ion traverses the gap between electrodes twice in each revolution and acquires an increment of energy on each passage of the gap which is given by the radio frequency potential between electrodes at that instant. In Fig. 2 this is illustrated on a voltage-time graph. Ions crossing at times labelled 1 will gain the maximum energy. Those having other phases relative to the oscillating electric field, such as at times 2–5, will gain energy at a slower rate and will need to make more revolutions to attain the same final energy. Ions can be accelerated during one half the cycle, and there will be electric focusing only within the quadrant indicated by the points 1, . . . , 5 As long as the magnetic field B remains constant, and the energy is small (so that the mass of the ion is essentially constant) this resonance will continue, with the particles increasing in energy and radius of path. At the periphery, radius R, the final kinetic energy for an ion of unit charge is given by:

$$T = \frac{1}{2}\frac{e}{m} B^2R^2 \tag{3}$$

Evaluating for protons, deuterons, and He^{++} this becomes

Protons: T (Mev) $= 3.12 \times 10^{-4}B^2R^2$
Deuterons: T (Mev) $= 1.56 \times 10^{-4}B^2R^2$
He^{++}: T (Mev) $= 3.12 \times 10^{-4}B^2R^2$ (due to the double charge)

where B is in kilogauss and R is in inches (units established by long usage in the cyclotron field).

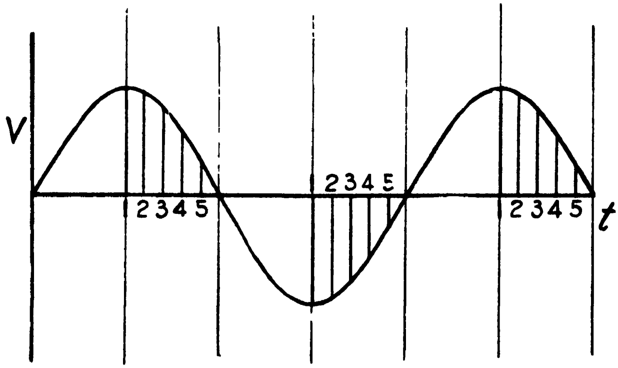


FIG. 2.—Voltage-time graph of the electric potential between the accelerating electrodes of a cyclotron. Resonant ions which cross the gap between electrodes at a phase 1, 2, . . . or 5 receive the same acceleration on each traversal of the gap. The final energy is the sum of the individual increments of energy.

As the energy and velocity of the ions increase, however, the mass increases in a relativistic manner, so that the e/m ratio for the ion is no longer a constant. Energy is associated with mass through the well known relativistic relations:

$$m = m_0(1 - \beta^2)^{-1/2}; \qquad \beta = \frac{v}{c} \tag{4}$$

$$E = mc^2 = E_0 + T = m_0c^2 + T \tag{5}$$

where m_0 is the rest mass, E the total energy, E_0 the rest energy, and T the kinetic energy. So, we can express eq. 2 in relativistic terms as:

$$f_i = \frac{ec^2B}{2\pi(E_0 + T)} \tag{6}$$

When the kinetic energy becomes large enough to be significant relative to the rest energy, the frequency of ion rotation will decrease. (The rest energy of a proton is 938 Mev and for the deuteron it is 1876 Mev; kinetic energies in excess of about 10 Mev will cause an appreciable change in resonant frequency.) As a consequence, the ions take longer to traverse a circular path than the fixed period of the oscillating electric field, and so they drift in phase until they cross the gap at a time when the voltage is zero. Since they are no longer being accelerated, this represents a limiting size and energy for the cyclotron.

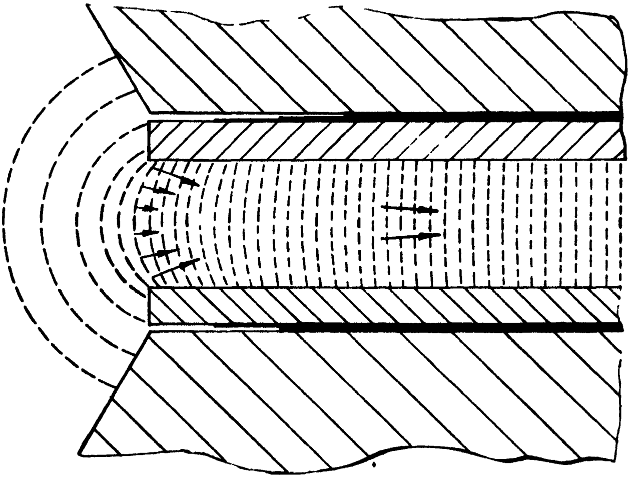


FIG. 3.—Focusing magnetic field between magnet poles in a cyclotron. Due to the radially decreasing magnetic field established by tapered shims in the central region and occurring naturally at the edge due to "fringing," the ions in a cyclotron experience magnetic forces which restore them to the central plane, resulting in an effective focusing of the ions.

A similar result occurs if the magnetic field decreases with radius; the resonant frequency decreases proportionately as shown in eq. (6). Such a radially decreasing magnetic field is found to be essential, however, in order to provide adequate magnetic focusing, and is accomplished by "shimming" the magnetic field. The lines of force produced by a radially decreasing field are concave inwards and so produce a vertical component of force on ions above or below the median plane, as illustrated in Fig. 3. This provides a strong focusing of ions towards the median plane. Conversely, a radially increasing magnetic field, which might be useful in canceling out the change in frequency due to increasing energy, will result in defocusing of the ions and an impractically small beam intensity.

Ions are produced at the center of the chamber in a small region between accelerating electrodes. The highest intensities are realized with a discharge tube electron source[9] using a heated cathode enclosed in a metallic cavity into which gas is fed. The several amperes of electrons emitted by such a discharge tube are collimated by the magnetic field into a vertical beam crossing between electrodes, where they ionize the gas also emerging from the source. The high frequency electric field between electrodes pulls the ions out of this region and into spiral paths in the median plane. In this early phase, when ion energies are low, the electric focusing during accelerations is more important than magnetic focusing. This can be illustrated qualitatively by reference to Fig. 4, which shows a schematic cross section of the electrodes and the path of an ion entering the gap off the median plane. The lines of electric field intensity are indicated, showing convergent forces on entering and a divergent force on leaving. However, since the ion has a higher velocity on leaving the gap, the time spent in the divergent field is shorter. Furthermore, if the oscillating electric field is decreasing during the time required to cross the gap, the magnitude of the defocusing force is smaller. The net result is a strong focussing toward the median plane during the quarter-cycle indicated in Fig. 2, for low velocity ions.

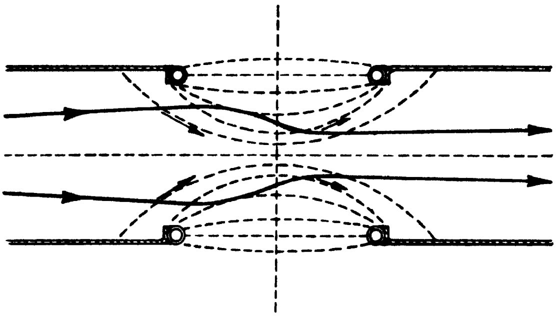
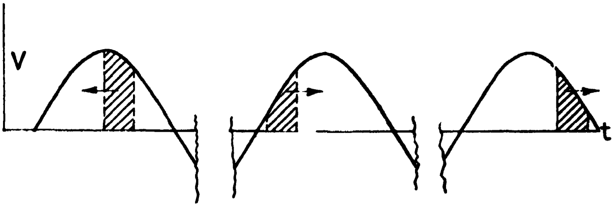


FIG. 4.—Electric focusing by cyclotron electrodes. Schematic diagram of cyclotron electrodes with an accelerating electric field indicated. Convergent forces on entry are larger than divergent forces on exit, resulting in a net focusing of the ions.

The limiting energy for the cyclotron can be shown to be a function of the number of revolutions, or the voltage between electrodes:[8]

$$T_m \simeq 1.8V^{\frac{1}{2}} \text{ (}V\text{ in kilovolts, for deuterons)} \qquad (7)$$

The larger cyclotrons[10] use high radio frequency power and high electrode voltages (100–150 kilovolts from each electrode to ground) to reduce the number of revolutions. A careful design of magnetic field is also required, shaped to give the smallest radial decrease compatible with focusing. In practice the electrode frequency is set to be slightly less than the initial ion frequency, yet greater than the final value. The result is a migration in phase, first backward, then forward on the graph of Fig. 1, with a total phase migration of less than π radians. This is illustrated in Fig. 5, where the band of ions which are being accelerated is indicated by shading in three representative radio frequency cycles, one at the start where maximum voltages are needed for high ion output from the source, one at an intermediate time when the phase shift has reached one extreme, and one at the final acceleration when the phase has swung to the other extreme.



a) FIRST ACCEL. b) INTERMEDIATE ACCEL. c) FINAL ACCEL.

FIG. 5.—Migration of phase of acceleration in the cyclotron.

(a) First acceleration with ions receiving near maximum acceleration where the electric field causes focusing.

(b) Intermediate acceleration where the phase has shifted to the extreme limit for acceleration and focusing is provided by the magnetic field.

(c) Final acceleration at the other extreme of phase shift.

Present 60-inch pole face diameter cyclotrons (200-tons weight) have achieved over 20 Mev deuterons by taking advantage of such techniques. It seems probable that 30 Mev represents a practical upper limit for fixed frequency cyclotrons. Using He^{++} ions the available energy is doubled, due to the double charge, and He^{++} ions have been produced with energies up to 40 Mev. Targets placed between the electrodes ("probe" targets) near the periphery will be bombarded with the total resonant ion currents; probe currents as large as 1 milliampere of 12 Mev deuterons have been observed.[8] The emergent beam deflected out of the chamber is divergent and usually has only 20–30% the intensity of the internal beam. Physical experiments with the high energy ions and volatile targets require such an emergent beam and a great deal of effort has been expended to perfect the technique and increase intensity. The greatest reported emergent beam is 420 microamperes of 12 Mev deuterons at St. Louis. Over 30 cyclotrons have been built in this country and abroad, so that by now a cyclotron is considered standard equipment for a well-equipped nuclear laboratory.

IV. Induction Accelerators: The Betatron

The cyclotron principle is not adaptable to the acceleration of electrons because of the rapid onset of the relativistic limit. The instrument which has been most successful for electrons is the induction accelerator or "betatron" proposed by Wideroe[5] in 1928, first operated successfully by Kerst[11] in 1939, and further developed at the research laboratories of the General Electric Co.[12] and the Allis-Chalmers Manufacturing Co.

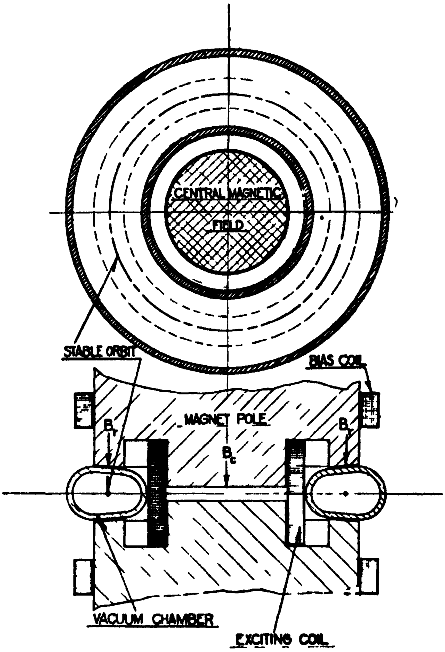


Fig. 6.—Betatron. Electrons are accelerated by induction due to changing central field B_c and travel in a circular orbit in the guide field B_r.

The betatron is a transformer in which a cloud of free electrons, located inside a doughnut-shaped vacuum chamber, takes the place of the secondary winding. The electrons move in a circular orbit of constant radius within the vacuum chamber as illustrated in Fig. 6. They gain energy by induction, by virtue of a changing magnetic flux Φ linking the orbit.

The induced voltage/turn is $\frac{d\Phi}{dt}$, as for a transformer, and the electric field (voltage/unit length) is given by:

$$\mathcal{E} = \frac{1}{2\pi r}\frac{d\Phi}{dt} \tag{8}$$

The force on the electron is $e\mathcal{E}$, and following the general law of motion can be expressed as rate of change of momentum:

$$\frac{d}{dt}(mv) = e\mathcal{E} = \frac{e}{2\pi r}\frac{d\Phi}{dt} \tag{9}$$

To maintain motion in a circular orbit of constant radius r, the magnetic field B at the orbit must increase as electron energy increases. By rearrangement of terms in eq. (1), we obtain:

$$mv = erB \tag{10}$$

So:

$$\frac{d}{dt}(mv) = er\frac{dB}{dt} \tag{11}$$

The condition for inductive acceleration at constant radius comes from equating the rate of change of momentum terms in eqs. (9) and (11):

$$\frac{d\Phi}{dt} = 2\pi r^2\frac{dB}{dt} = 2\frac{d}{dt}(\pi r^2 B) \tag{12}$$

The betatron relation derived above says that in any increment of time the linking flux Φ must change at a rate *twice* that which would occur if the central magnetic field were uniform and equal to the field at the orbit. The "2:1 rule" holds for relativistic energies as well as in the nonrelativistic range, since it was derived by treating the momentum mv as a single variable. This flux relation requires a central iron core with high flux density inside the orbit. Since the induced voltage is determined by the rate of change of flux, the iron core is laminated as in a transformer, and alternating power at 60 or 180 cycles is used to produce the changing magnetic field. A time plot of an acceleration cycle is illustrated in Fig. 7 for 60-cycle repetition rate.

Electrons are injected into the vacuum chamber from a "gun" in which thermionic electrons from a hot cathode are accelerated and focused by a potential of several kilovolts. The slit is located at the outer rim of the vacuum chamber and directs the electrons into their first circular path. Due to the oscillations set up about the equilibrium orbit a small fraction of the injected electrons miss the walls of the chamber and the back of the gun on subsequent revolutions and are captured in stable betatron orbits. The injection is timed so that the magnetic field at the orbit is correct for the injection energy.

During the first quarter-cycle, as the magnetic field builds up from zero, the electrons are accelerated through hundred of thousands of revolutions, gaining only a few hundred volts per turn by induction. To prevent deceleration during the succeeding quarter-cycle, the electrons are deliberately thrown out of position as they approach peak energy by means of an auxiliary coil which distorts the magnetic field so that the electron orbit is shifted radially or axially to strike a target. This produces a beam of x-rays from the target which is sharply collimated in the direction of the impinging electrons. The x-rays, having a continuous spectrum with maximum energy equal to the electron energy, emerge from the tube through the vacuum wall, where they are available for experiments.

Oscillations of the electrons about the theoretical or "equilibrium" orbit will occur if the electrons are displaced from this orbit. These oscillations will be of two types: vertical and radial. Stable oscillations will result if restoring forces exist which oppose a deviation from the equilibrium orbit. For vertical deviations there must be vertical restoring forces, and for radial displacements, radial restoring forces. Deviations of the ions from the median plane can be limited and controlled by a "focusing" magnetic field, as illustrated in Fig. 3. To provide such focusing the field near the orbit must decrease with increasing radius, as in the cyclotron, usually accomplished by tapering the gap between magnet poles. The degree of tapering affects the radial oscillations, which will be stable if the field does not decrease with radius more rapidly than $1/r$. A complete analysis of these betatron oscillations showing the effect of magnetic field tapering and the way such oscillations are damped by the increasing magnetic field has been published.[13,14] The oscillations will be considered in more detail in a later section.

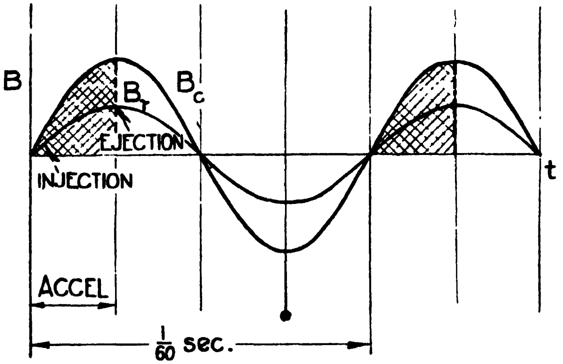


Fig. 7.—Betatron acceleration cycle. Electrons are accelerated during the first quarter-cycle where the central field B_c and the guide field B_r are increasing proportionately.

A considerable saving in weight of iron is obtained with D.C.-biasing of the central, accelerating magnetic field. Additional exciting coils powered by D.C. are used to bias the central field so that it goes from a large negative to a large positive value during the time in which the deflecting field at the orbit goes from zero to its final value. This comes from the relation of eq. (12) which requires only a proper positive value of $\frac{d}{dt}(\Phi)$, not necessarily a positive Φ. The principle is illustrated in Fig. 8. In this manner the particles are accelerated by a larger total change in Φ, and (for the same weight of iron) to a higher energy.

Power to drive the transformerlike magnet can be kept to a minimum by resonating the inductance with a condenser bank at the desired frequency. In this way circulating power to supply the large amount of stored energy in the magnetic field can be obtained with a relatively small power input. The power absorbed will be just that dissipated in heat losses in the copper windings, the iron core, and the condensers.

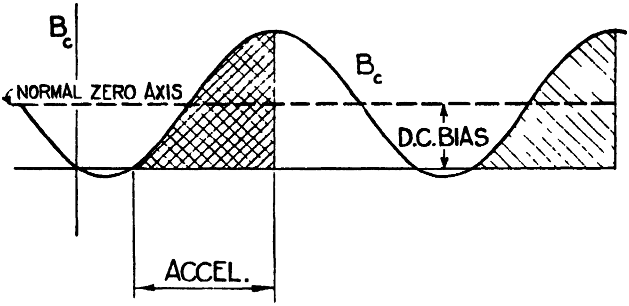


FIG. 8.—Biased betatron accelerating cycle. A D C. bias applied to the central field increases the magnitude of the flux change and results in higher electron energy.

Betatrons of several sizes have been constructed and operated. The original Kerst machines produced electrons of 2.3 and 20 Mev. The largest in operation is the 100 Mev machine at the General Electric Research Laboratory.[12] The largest now planned is being designed by Kerst at the University of Illinois, and is expected to reach 300 Mev.

V. PRINCIPLES OF ACCELERATION TO HIGH ENERGIES

The brief descriptions above of direct voltage accelerators, the cyclotron, and the betatron are sufficient to show the status of accelerator design before the war. These machines had been developed essentially up to their energy ceilings. They are relatively simple applications of the equations of motion and do not utilize fully all of the possibilities. As an introduction to the several new types of accelerators, which have promise of much higher energies, it seems proper to present and analyze the equations of motion of charged particles, and to describe the properties of motion which lead to phase stability.

1. Relativistic Equations of Motion

The same fundamental equations of motion apply to all accelerators and to all charged particles, and must be expressed in relativistic terms. Only two fundamentally different geometrical arrangements are possible: that for motion in a straight line and that for circular motion in a magnetic field. The equation of motion for linear acceleration can be expressed simply as:

$$\frac{d}{dt}(m\dot{x}) = \sum F(x, t) \tag{16}$$

Here m is the relativistic mass and the ΣF term represents a complicated function of position and time which describes the total electrical accelerating force in the direction of motion. In the modern linear accelerator this usually involves multiple electrode structures and electric fields arranged to synchronize in time with the motion of the particles. The elementary equation above does not describe the transverse forces which are important for focusing. A solution to this linear equation can usually be obtained if the ΣF term can be expressed analytically. Such a solution will show the special conditions of frequency and phase of the electric field which will result in acceleration.

The betatron, synchrotron, cyclotron, frequency modulated cyclotron, and frequency modulated synchrotron all use magnetic fields to produce motion in a circle and the dynamical equations of motion are fundamentally identical. For circular motion in an axially symmetric magnetic field, with external torques applied about the axis of symmetry, cylindrical coordinates give the simplest formulation of the equations of motion. Neglecting second order terms, these can be written in the form:

$$\frac{d}{dt}(m\dot{r}) = mr\dot{\theta}^2 - er\dot{\theta}B_z \tag{17a}$$

$$\frac{d}{dt}\left(mr^2\dot{\theta} - \frac{e\Phi}{2\pi}\right) = \frac{dw}{d\theta} - \frac{dL}{d\theta} \tag{17b}$$

$$\frac{d}{dt}(m\dot{z}) = er\dot{\theta}B_r \tag{17c}$$

In the above equations m.k.s. rationalized units are used:

m is the relativistic mass of the particle;

B_z and B_r are the vertical (axial) and radial components of the magnetic field at the orbit. The time rate of change of magnetic field will be slow, so that curl $B = 0$ and $\frac{\partial B_r}{\partial z} = \frac{\partial B_z}{\partial r}$ in the region of the orbit.

Φ is the magnetic flux linking a circular orbit of radius r;

$\frac{dw}{d\theta}$ is the external electrical torque about the axis of symmetry where dw is the work done by this torque during a displacement $d\theta$;

$\frac{dL}{d\theta}$ is the decelerating torque associated with energy radiated by the particle due to its motion in a circular path. These equations apply directly for electrons, and can be adapted to positively charged particles by changing the sign of e.

Completely general solutions to these equations of motion would be unnecessarily complicated and of questionable value. Specific solutions for the special cases representing individual machines are of more practical interest. For example, the relation derived for the cyclotron as eq. (1) for motion in a circular path at constant radius, comes directly from eq. (17a) by inserting the condition; $\dot{r} = 0$:

$$\dot{\theta} = \omega = \frac{v}{r} = \frac{eB_z}{m} \tag{18}$$

Orbit radius R for a given kinetic energy and magnetic field can be obtained by use of the well known relativistic relation:

$$(mv) = \frac{(E^2 - E_0{}^2)^{\frac{1}{2}}}{c} \tag{19}$$

Substituting in eq. (18) we obtain:

$$R = \frac{mv}{eB} = \frac{(E^2 - E_0{}^2)^{\frac{1}{2}}}{ceB} = \frac{[T(T + 2E_0)]^{\frac{1}{2}}}{ceB} \tag{20}$$

For T and E_0 in Mev units and B in webers/m.^{2} (units of 10,000 gauss) this becomes:

$$R \text{ (meters)} = \frac{1 \times 10^{-2}}{3B} [T(T + 2E_0)]^{\frac{1}{2}} \tag{20a}$$

This relation is plotted in Fig. 9 for electrons, protons, deuterons, and alpha particles of energies up to 10^{10} electron volts and for magnetic fields up to 10 webers/m.^{2} (To use the chart pick a maximum energy T on the horizontal coordinate, find the BR for the particle involved on the left hand ordinate, follow this BR value to the intersection with the chosen value of B and read the radius R on the upper scale.) In Table I a few typical values are listed to show the dimentional requirements of magnetic accelerators. Note the approach to a linear relation between radius and energy for relativistic energies where $T \gg E_0$. Note also the converging dimensions for electron and proton accelerators at very high energies.

TABLE I. Orbit radius in meters at 1 weber/m.² (10,000 gauss).

T (Mev)	Electrons	Protons	Deuterons	He⁺⁺ ions
1	0 00485	0 143	0 206	0 144
10	0 0347	0 454	0 643	0 454
100	0 330	1 47	2 05	1.45
1000	3 30	5 60	7 22	4 85
10000	33 0	36 0	38 7	21 8

The kinetic energy of the particles in terms of orbit radius and magnetic field comes from a rearrangement of eq. (19):

$$T^2 + 2TE_0 = c^2e^2B^2R^2 \tag{20}$$

This is in the form of a quadratic in T, but can be read off the plot of Fig. 9 by reversing the above procedure.

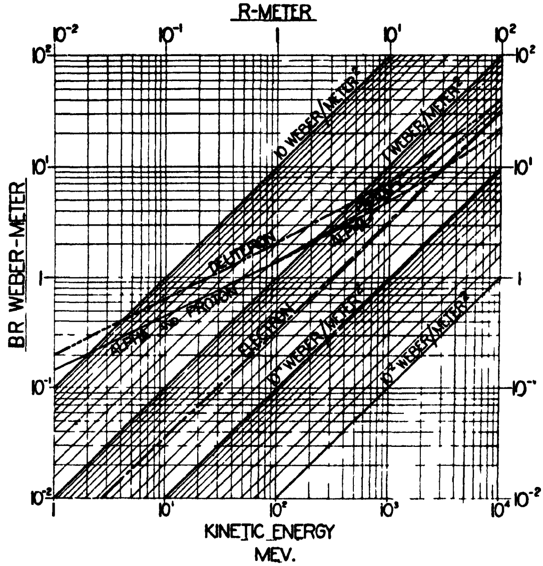


FIG. 9.—Energy vs. radius in a magnetic field. For a particle energy in Mev on the lower scale find the corresponding BR on the vertical scale using the appropriate particle curve. Follow this value of BR to the desired magnetic field in webers/m. and read orbit radius in meters on the upper scale.

Ion rotation frequency, $f_i = \frac{\omega}{2\pi}$, requires another rearrangement of eq. (18):

$$f_i = \frac{eB}{2\pi m} = \frac{c^2eB}{2\pi(E_0 + T)} \tag{21}$$

or, in terms of radius:

$$f_r = \frac{v}{2\pi R} = \frac{c}{2\pi R}\beta = \frac{c}{2\pi R}\left[1 - \left(\frac{E_0}{T + E_0}\right)^2\right]^{\frac{1}{2}} \tag{22}$$

We will make use of this relation later when considering the frequencies required for the several accelerators. The resonance frequency for the acceleration of particles of different energies is plotted in Fig. 10. (To use this chart start with the energy T and find the value of the quantity $\frac{\omega R}{c}$ for the particle concerned on the left hand ordinate, then follow this value to the intersection with the chosen radius and read the resulting linear frequency on the upper scale) Note the linear dependence of f_r, ω and β.

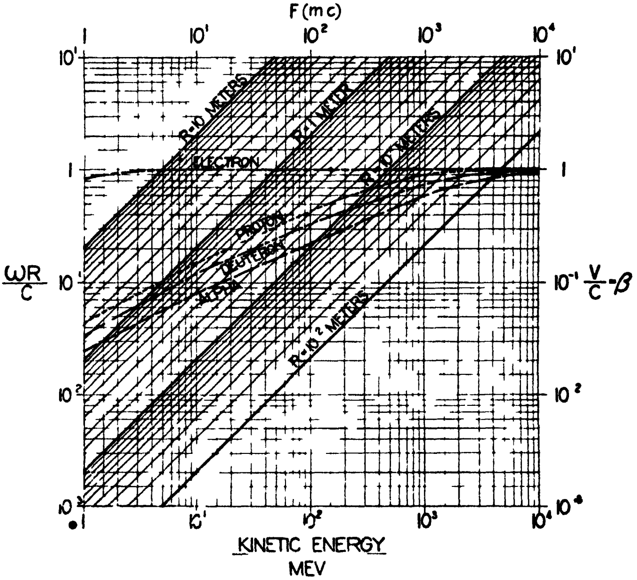


FIG 10—Energy vs frequency at known orbit radius For the chosen particle and an energy value given on the lower scale find the appropriate value of $\beta = \frac{\omega R}{c}$ on the vertical scale Follow this value of β to the intersection with the orbit radius in meters and find the frequency in Mc on the upper scale

Eqs. (17a) and (17c) show the variations in the radial and axial coordinates. Present "good" magnet designs are symmetrical about the median plane and are arranged to have radial components of magnetic field B_r off this plane which result in focusing in the z coordinate. Such focusing forces will restore the particles to the orbit and will set up oscillations about the true orbit. Qualitatively the field should decrease with increasing radius, as illustrated in Fig. 3. Quantitatively, the rate of change of B with radius should be so adjusted that the radial and vertical oscillations are stable and their amplitudes are small. The index n which is used to define this shaping of the field near the orbit is given by:

$$B_r = B_0 \left(\frac{r_0}{r}\right)^n \tag{23}$$

where B_0 is the field at the orbit position r_0 and B is the field at a slightly different radius r. By differentiation we obtain:

$$n = -\frac{r}{B}\frac{dB}{dr} \cdots \tag{24}$$

A radially decreasing magnetic field having any positive value of n will provide stability for vertical (z) oscillations and the amplitude will decrease with increasing values of n. Radial oscillations will be stable if the field does not decrease with radius more rapidly than $1/r$ or in other terms, for $n < 1$. To understand this, note that the magnetic force is equal to the "centripetal force" at the equilibrium radius, or in terms of eq. (17a):

$$er\dot{\theta}B_z = mr\dot{\theta}^2 \text{ (at equilibrium radius)} \tag{25}$$

If the particle is to be restored to the equilibrium orbit when displaced radially, the magnetic force should be larger than the centripetal force for larger radii, and for smaller radii the converse should hold. It can be shown[14] that this is true if the value of n does not exceed unity. So, the range of values for stability of both radial and vertical oscillations is: $0 < n < 1$. The magnetic field index n is one of the fundamental parameters in the design of accelerators and will recur in the discussions to follow of the several types.

The frequency of the oscillations is found to be a simple function of n, given by:

$$\omega_z = \sqrt{n}\,\omega \tag{26a}$$

$$\omega_r = \sqrt{1-n}\,\omega \tag{26b}$$

Here ω is the particle rotation frequency and ω_z and ω_r are the frequencies of the vertical and radial oscillations respectively. A simple visualization of the free oscillations will help to justify these conclusions. Consider, for example, a particle of energy T in a uniform magnetic field B such that the true orbit radius is R. Consider, furthermore, that this particle is displaced outward to a position $R + x$. It will describe a circle of radius R, crossing the true orbit twice/revolution and having a radial position $R - x$ at the far side of the orbit, which represents a radial oscillation of frequency ω and amplitude x. Since $n = 0$ for a uniform field this is in agreement with eq. (26b). Under the same conditions a particle displaced from the median plane vertically (say by $+z$) will rotate in an orbit parallel to the median plane and never cross it, representing an oscillation frequency $\omega_z = 0$ as indicated in eq. (26a). For other allowed values of n the oscillation frequencies will be smaller than ω, but still of the same order of magnitude. At $n = 0.5$, for example, $\omega_z = \omega_r = 0.707\omega$. These free oscillations are also known as betatron oscillations; it was through the prediction of these oscillations and utilization of the resultant focusing that Kerst was able to make a practical betatron from the qualitative ideas of Wideröe.

For large values of n (near unity) the radial component of magnetic field B_r is large, the focusing effect on the vertical oscillations is a maximum, and the vertical amplitudes are small. On the other hand, the field decreases considerably with radius so the radial excursions will be large. For small values of n (near zero) the reverse is true; radial oscillation amplitudes will be a minimum and vertical amplitudes a maximum. The oscillation amplitudes are equal for $n = 0.5$. Oscillation amplitudes will also vary with the intensity of the magnetic field, decreasing as the field (and particle energy) increases:

$$A \sim B_z^{-\frac{1}{2}} \tag{27}$$

This damping will result in the particles being focused tightly around the true orbit at high energies, a useful practical result.

A qualitative interpretation of the energy relations will help to show the significance of the several terms in eq. (17b). For example, total particle energy can be obtained by multiplying eqs. (17a), (17b), and (17c), by $\dot{r}$, $\dot{\theta}$, and $\dot{z}$ respectively, and adding. We obtain:

$$\frac{d}{dt}(mv^2) = \left(\frac{e}{2\pi}\frac{d\Phi}{dt} + \frac{dw}{d\theta} - \frac{dL}{d\theta}\right)\frac{d\theta}{dt} \tag{28}$$

In this form the total linear kinetic energy mv^2, is shown to depend on three energy terms:

Firstly, the flux Φ enclosed within an orbit of radius r is given by:

$$\Phi = 2\pi \int_0^r rB_z dr \tag{29}$$

Time rate of change of flux, $d\Phi/dt$, determines the induced voltage applied to the particles. This is the well known transformer principle responsible for acceleration of electrons in the betatron. The "2:1" rule comes directly from the first two terms of eq. (17b) by neglecting the two right hand terms. Induction can be neglected for the cyclotron where B_z is constant, but must be included in the calculations for any machine in which magnetic field varies with time, such as the synchrotron.

Secondly, the electrical torque $dw/d\theta$ is usually applied by means of one or more accelerating gaps between high frequency electrodes, as in the cyclotron. However, the number of particle revolutions is generally so large (the order of 10^4 to 10^6), and the fractional increase in energy/revolution so small, that we can treat the acceleration process adiabatically and describe the torque in terms of an azimuthally uniform accelerating torque which varies slowly with time. So electrical torque can be expressed as:

$$\frac{dw}{d\theta} = \frac{Ve}{2\pi} \tag{30}$$

where V is the potential difference traversed/*per* turn by the particle of charge e.

Thirdly, the radiation loss term $dL/d\theta$ depends upon the radial acceleration of the particle. Blewett[15] has derived a relation for the energy radiated per revolution:

$$\Delta L = \frac{4\pi}{3} \frac{e^2}{R} \left(\frac{E}{E_0}\right)^4 \tag{31}$$

This factor is significant only for electrons with energies above about 200 Mev. The rapid increase of radiation loss with energy spoils the betatron balance and will require excessive electrical accelerations in the synchrotron. It may well be the energy limiting factor in magnetic electron accelerators, but due to the much larger rest energy of protons it can be neglected for proton accelerators.

So we see that particles in a magnetic machine can be accelerated either by "betatron" induction or by externally applied electric fields. With induction acceleration there is no critical frequency involved, but external electric accelerating fields must be oscillatory in character, and adjusted in frequency to synchronize with the rotational frequency of the charged particles. If the particles are electrons with energies above a few Mev, the velocity is essentially equal to the velocity of light and the frequency essentially constant, as shown by eq. 22 and Fig. 10. The much heavier protons will undergo continuous relativistic changes in mass and so the frequency will vary widely in the energy region between about 10 Mev and 10 Bev.

We will leave the equations of motion at this point with these qualitative interpretations, which are sufficient, however, to point out the basic requirements of superenergy accelerators.

2. *Principles of Phase Stability*

Multiple accelerations will be necessary to achieve superenergies without exceeding the voltage/acceleration set by the practical limits of insulation and power in machine construction. In resonance-type accelerators the primary problem is that of keeping the particles in step with the accelerating electric field for a very large number of accelerations. In the cyclotron resonance is maintained only so long as the natural frequency of rotation of ions is constant, and breaks down when this frequency begins to decrease owing to the relativistic increase in mass as ions reach high energies. The number of accelerations is limited to only a few hundred, and becomes smaller as maximum energy increases. In the linear accelerator a similar limit exists, set by the precision of the applied frequency in matching the precut electrode spacings.

A new impetus was given to the design of high energy accelerators with the announcement in 1945 by McMillan[16] and (independently) by Veksler[17] of the principle of phase stable orbits in magnetic resonance accelerators. Proposals were made for machines utilizing this principle, and the name "synchrotron" was suggested by McMillan. This principle, in brief, involves the concept of stable circulating orbits for charged particles moving in a cyclotron-type accelerator, in which the particles increase in energy as a result of a slow variation of the magnetic field, of the frequency of the accelerating electric field, or of both. An immediate application was to the 184-inch Berkeley cyclotron which was arranged as a phase stable accelerator by applying a modulated frequency to the electrodes.[18] The success of this "synchro-cyclotron" or "frequency modulated cyclotron" has afforded convincing proof of the validity of the concept of stable, synchronous orbits for light positive ions in a cyclotron. Shortly after, two electron synchrotrons[19,20] were successfully operated in which the magnetic field was the variable and the frequency was constant, again demonstrating the validity of the method. Now several electron synchrotrons and synchro-cyclotrons for energies up to 300 Mev are under construction, and designs have been proposed for even higher energies.

The phase stability inherent in the synchrotron can best be understood by a qualitative physical description of the motion of the charged particles. Consider, for example, the "stationary" orbits possible for a particle moving in a circular path in a uniform magnetic field, on each revolution crossing a gap between accelerating electrodes on which is applied an oscillating electric field with a frequency identical with the frequency of rotation of the particle. As a start consider those particles which cross the gap at instants of time when the oscillating electric field is crossing zero (zero phase). This is illustrated by the points labeled 0, 2π, 4π, etc. on the voltage-time graph of the electric field of Fig. 11. These particles will neither gain nor lose energy and will continue to rotate at constant frequency and in the same orbit. Now to see that this orbit is stationary, consider a particle which crosses the gap at an earlier time (positive phase) such as at t_1. It will gain energy, will rotate at a slightly lower frequency as shown by eq. (21), and will take a somewhat longer time to return to the gap, as illustrated by points t_2, t_3, etc. in subsequent accelerations, when the voltage across the gap is smaller. Eventually the particle will cross the gap at zero phase, but with an accumulated excess of energy so that it will continue the phase shift into the decelerating part of the cycle. Now the situation is reversed and the particle loses energy; its frequency increases and it is returned to the zero phase position again. This represents a phase oscillation about the equilibrium phase, $\Phi = 0$. Note that the alternate situation at $\Phi = \pi$, 3π, etc. will result in instability.

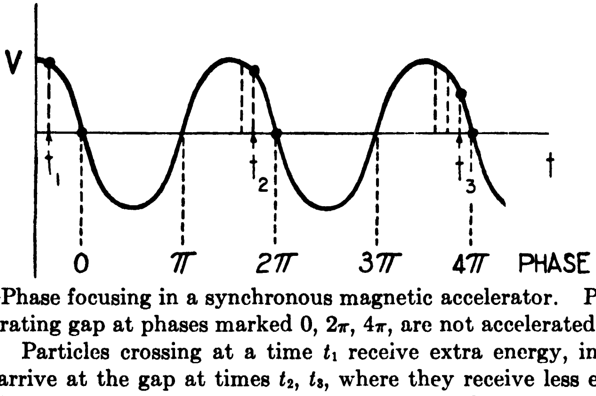


FIG. 11.—Phase focusing in a synchronous magnetic accelerator. Particles crossing the accelerating gap at phases marked 0, 2π, 4π, are not accelerated and continue in resonance. Particles crossing at a time t_1 receive extra energy, increase in frequency, and arrive at the gap at times t_2, t_3, where they receive less energy. Continuation of this process results in phase oscillations around zero phase.

The particle considered above will traverse orbits of different radius as the energy varies, in accordance with eq. 20. As long as the particle energy exceeds the equilibrium energy it will travel in orbits of larger radius and conversely it will make smaller circles for energies below the normal value. This "breathing" motion of the orbit represents a radial oscillation associated with the phase oscillation described above, and in addition to any radial free oscillation.

Now let us consider particles which have initial energies differing slightly from the equilibrium energy. Such particles will be travelling in orbits which are larger or smaller than the equilibrium orbit and so will have rotational frequencies which differ from the applied electric field. By an argument similar to that above it can be show that these particles will also oscillate about the equilibrium energy and the equilibrium radius, with the oscillations slightly displaced in phase from those illustrated in Fig. 11.

When a small change is made in the magnetic field, or the applied electric frequency, the particles will temporarily experience a phase shift, and will adjust themselves to the new condition by gaining or losing sufficient energy to re-establish equilibrium conditions. *Furthermore, if either the frequency is decreased, or the magnetic field is increased slowly and continuously, the particles will follow this change and gain energy at a steady rate determined by the rate of frequency or magnetic field modulation.* From these two alternatives come the applications known as the synchrotron and the synchro-cyclotron. The average energy gained per acceleration under these conditions determines an equilibrium phase Φ_e, about which the phase oscillations are centered. This is illustrated in Fig. 12a in which the phase amplitude is shown as extending between arbitrary limits Φ_1 and Φ_2 of stable oscillations. In the extreme case a particle will migrate in phase between these limits, making several hundred revolutions to complete the phase oscillation. Particles which deviate less from equilibrium conditions will perform oscillations of smaller amplitude and with slightly different phase oscillation frequencies.

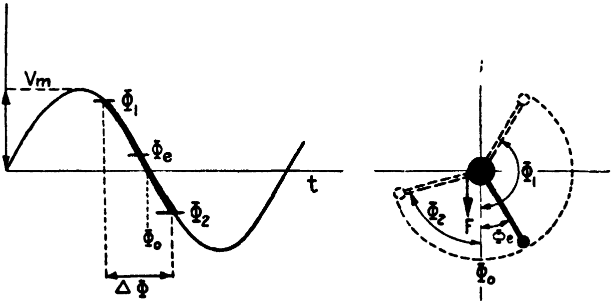


FIG. 12.—(a) Phase oscillation range $\Delta\Phi$ about an equilibrium phase Φe illustrated on voltage-time plot of accelerating electrode. (b) Pendulum with constant torque analogue of phase oscillations, showing the equilibrium displacement Φe and the region of stable oscillations Φ_1 to Φ_2.

Several authors of papers on the theory of the synchrotron have pointed out the analogy of these phase oscillations to the motion of a physical pendulum to which a constant torque is applied. The equations of motions are identical if the angular displacement of the pendulum is substituted for the phase angle in the synchrotron. In Figure 12b such a pendulum is presented schematically so that the similarity of the motion to that of the phase oscillations can be visualized. The pendulum has a region of stable oscillations about the equilibrium position Φ_e

extending from $\pi - 2\Phi_e$ at one extreme to $-(\pi - 2\Phi_e)$ at the other. If the amplitude exceeds this limit the pendulum will cease to oscillate and will spin continuously in one direction. With the synchrotron the particle which exceeds the limits will gain or lose too much energy for stability and so will spiral inward or outward until it hits the walls of the synchrotron chamber.

Another visualization of phase oscillations is illustrated in Fig. 13. Here the oscillations in energy of a particle are portrayed as a function of time, with the equilibrium energy T_e increasing slowly. This can be interpreted as a potential energy diagram with stable potential valleys in which particles are trapped. The dots represent consecutive positions of a particle which is oscillating within this potential valley. The limit of phase oscillation appears here as the peak of a potential hill, Φ_1. As the oscillations progress, the pattern will progress upward and to the right along a line which represents the average energy increase, carrying the trapped particles to higher energies. If a particle crosses over the rim of one trough, it will be lost from synchronism.

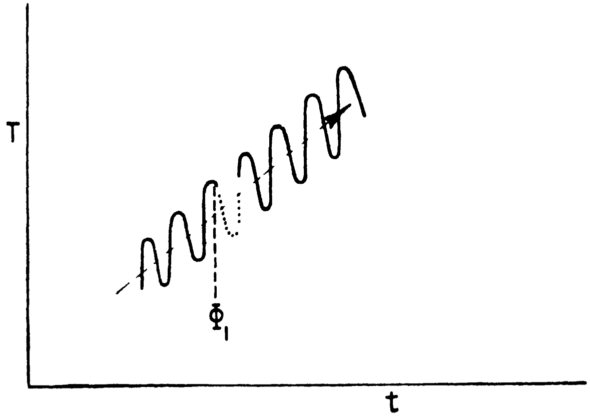


FIG. 13.—Energy oscillations in a synchronous accelerator. Ions are trapped in a potential valley and oscillate as indicated by the dots representing consecutive revolutions, limited by the phase Φ_1. The pattern progresses upward along a line representing the average increase in energy with time.

The discussion so far has shown that the synchrotron orbits are stable but has not described the focusing in the vertical or axial coordinate, nor has it demonstrated any damping of the amplitude of phase oscillations, both of which are essential to the practical success of the synchrotron principle.

To obtain focussing in the z coordinate the magnetic field must decrease slightly with increasing radius, exactly as described earlier for the betatron. Radial free oscillations will also be present, superimposed on the low frequency radial synchrotron oscillations. For stability of the free oscillations the magnetic index must be in the range: $0 < n < 1$. Damping will occur as the restoring forces increase in magnitude, if the magnetic field increases with time; with constant magnetic field the amplitude will remain constant. Synchrotron phase oscillations will not only be damped by an increasing magnetic field, but also by other parameters. Theoretical studies show that the rate of damping can be affected by the time rate of change of frequency or by the rate of change of potential across the accelerating gap. For example, an increasing peak voltage relative to the average voltage/turn (see Fig. 12) will increase the rate of change of potential while the particle is crossing the gap, and will cause a strong damping of phase oscillations. The conditions for phase oscillation damping are somewhat different for the different types of accelerators, and will be discussed in more detail for the specific machines.

Let us recapitulate by describing the expected shape of the bunch of particles in a phase stable orbit of a synchrotron. First, they will be distributed in phase about the equilibrium value, with a large initial phase amplitude or phase acceptance angle which depends primarily on the ratio of peak voltage to the required average voltage/turn, and will be damped slowly to smaller amplitude. This represents a distribution in time of crossing the accelerating gap, or an azimuthal distribution in angular position around the orbit. Next is the spread in radial location due to the associated oscillation around the equilibrium energy. The initial radial oscillation amplitude will be limited by the initial phase acceptance angle which defines the energy spread and by parameters such as the n value, and will be rapidly damped into a narrow band concentrated about the equilibrium orbit of the instant. This elongated bunch of resonant particles will rotate at the frequency of the applied electric field. The particle envelope is illustrated in Fig. 14 for a time near the start of the modulation cycle when energy is low and oscillations are large. Such phase and radial oscillations will have frequencies hundreds of times smaller than the particle rotation frequency.

Superimposed on the slow synchrotron oscillations will be the high frequency free oscillations identical with those in the betatron. These, are due to spatial deviations from the smoother orbits described above, such as would result from a radial location incompatible with the energy and magnetic field. The force which causes these oscillations about the true orbit is due to the focusing properties of the radially decreasing magnetic field near the orbit. As the magnetic field increases the amplitude of the oscillations is diminished. The frequencies of these free oscillations are smaller than, but of the same order of magnitude as, the rotation frequency. So these high frequency oscillations will be added to the smoother synchrotron orbits; this is illustrated qualitatively (but not to scale) by the scallops shown about the periphery of the particle envelope in Fig. 14. These add to the radial width of the envelope and establish the vertical dimension. Finally, this "sausage" shaped envelope will be damped to much smaller dimensions as particle energy increases, ultimately acquiring a long narrow shape, which depends in detail on design parameters of the synchrotron such as the magnetic index *n* and the ratio of peak voltage to average voltage/acceleration on the gap.

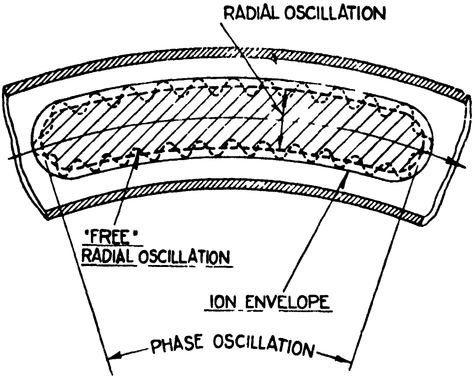


FIG. 14.—Schematic illustration of phase and radial oscillations in the synchrotron and the superimposed "free" or betatron oscillations, shown as an envelope of ion paths located in the accelerating chamber.

VI. THE SYNCHROTRON

The synchrotron accelerates electrons around an orbit of essentially constant radius and so requires only a narrow annular magnetic field such as can be produced by a ring-shaped magnet. It is this economy of magnet weight that provides the primary advantage over the betatron which needs in addition a large central flux core, or the cyclotron which uses a solid core magnet. Electrons are "injected" into this orbit at low energy and low magnetic field. An accelerating high frequency electric field is located at one point on the orbit and supplied by an external power source. The magnetic field is modulated in time at relatively low frequency (60 cycles/second). As the magnetic field increases the electrons are bunched and focused by phase oscillations so that they acquire energy at the proper rate to maintain an orbit of constant radius. On reaching maximum energy they are deflected against a target as in the betatron, in a short pulse. The cyclical operation of the magnet results in a corresponding pulse repetition frequency of 60 cycles/second.

A great deal of study has gone into the theory of synchrotron oscillations. Following the original articles by McMillan[16] and Veksler,[17] theoretical studies have been published by Dennison and Berlin,[21] Bohm and Foldy,[22] and Frank[23] which extend the analysis into the detailed dynamics of motion. All these studies are in basic agreement, varying only in the type of approximations used in the calculations and in the emphasis placed on the several aspects of the motion. They all find the same conditions for oscillation stability and for damping of oscillations. The theoretical understanding of the synchrotron seems to be adequate.

The distinguishing feature of the electron synchrotron is the essentially constant frequency of ion rotation above about 2 Mev energy, and the opportunity to use a fixed frequency accelerating field. Most designs use a high Q resonant circuit at one point in the "doughnut" vacuum chamber, such as a tuned cavity resonator. The basic frequency is determined by the angular velocity of electrons of velocity c. At 1 meter radius, for example, the frequency is:

$$f = \frac{c}{2\pi R} = 48 \text{ Mc}$$

It should be noted that harmonic frequencies will also produce resonance, such as 96 Mc in the example above. In this case the ions will be accelerated in two bunches instead of a single one, and each bunch will be correspondingly shorter.

The magnetic field index n must be in the range, $0 < n < 1$, as in the betatron. Most designers use a large value of n ($.5 < n < 1$) so that vertical oscillation amplitudes are less than the radial amplitudes; a choice of $n = \frac{2}{3}$ is common. At high energies, where radiation losses become significant, it has been shown that a value of $n = \frac{3}{4}$ leads to a pronounced antidamping, so this represents an upper limit. The spatial region between poles in which the magnetic field index is within these limits defines the useful aperture for electrons. Detailed model experiments, to determine the best tapering and shaping of pole faces to give the largest region of uniform n-values, have been an essential feature of all design studies. Several ingenious devices have been suggested to increase the aperture at the start of the cycle, when the greatest amplitudes occur. One of these is to add "lips" of high permeability ferromagnetic material at the edges of the gap to widen the aperture at low fields; these lips saturate at high inductions and so allow the fields to concentrate in the central region.

To power the magnet large alternating currents are needed at 60 cycle frequency. The most satisfactory method of supplying these currents is to resonate the inductance of the magnet with a capacitor bank. In this way the average power demand is reduced to the amount needed to supply heat losses in the copper and iron. In the General Electric Co. design[20] the condenser bank has a capacity of 16 microfarads rated for 21,000 volts; the power rating is 3000 kilovolt-amperes at 60 cycles.

Most synchrotrons use a "betatron start" to accelerate electrons by induction from about 50 kilovolts injection energy to about 2 Mev. At the General Electric Co. and the University of California this is done with a magnet similar to the betatron magnets, which has an internal flux core and a combined external magnetic return path. In the designs of The Massachusetts Institute of Technology and at some other laboratories a large number of "C" shaped magnet sections with individual external return paths are mounted around a circular orbit and the betatron flux is carried by "flux bars" of modest dimensions made of high permeability metal which are located inside the orbit. These flux bars "short" the magnetic field at low inductions but become saturated at high inductions. In Fig. 15 a sketch of this ring magnet synchrotron is shown to illustrate the principle of the machine and the saving in magnet iron possible with the synchrotron. Such magnets must be laminated

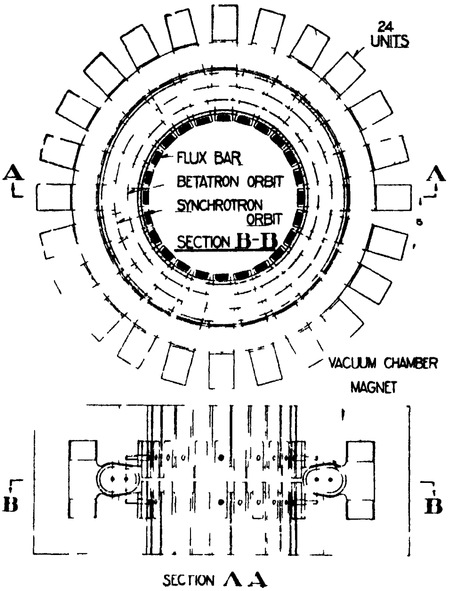


FIG 15—Electron synchrotron This sketch illustrates a ring magnet formed of "C" sections with internal flux bars to provide flux for a betatron start.

to reduce eddy current losses, and the design includes auxiliary clamps and mechanical structures which are omitted from the sketch for clarity.

The transition from betatron to synchrotron action must be made in such a way that the electrons "lock-in" to synchronism with the oscillatory electric field.[23] In the betatron region the particles are uniformly distributed around the orbit and are moving with an angular frequency ω given by eq. 18. If the betatron relation is obeyed they have essentially uniform energies and move at constant radius, except as perturbed by the radial and vertical oscillations. After the transition to synchrotron motion the electrons will be bunched in phase and azimuth about the equilibrium phase, at the driving frequency ω_1 (correct only for electrons of velocity c), and will be performing wide oscillations in energy and radius until damped by increasing energy and magnetic field. The "slip" frequency $(\omega_1 - \omega)$ must be small for "locking-in" to occur, and must be reduced to zero in a relatively short time, the order of a few hundred revolutions. This can be visualized as the transition from rotary to oscillatory motion in the pendulum motion described in section V*b*. Theory predicts[23] and experimental results show that this transition does take place smoothly and that a large fraction of the electrons in the betatron orbits are captured in synchrotron orbits.

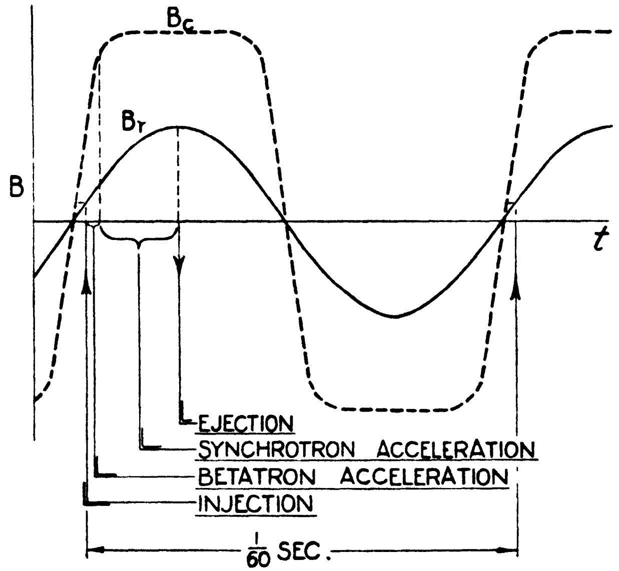


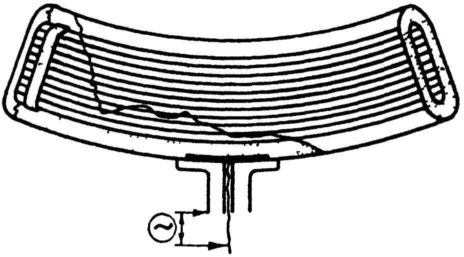
FIG. 16.—Acceleration cycle for a synchrotron with betatron start, showing 60 cycle guide field B_r and the saturating field in the flux bars B_c.

The time cycle for acceleration is illustrated in Fig. 16, in which the guiding field at the orbit B_r is produced by 60 cycle alternating current, and the field in the flux bars B_c rises to a maximum in the short interval labeled "betatron acceleration" and then holds this saturation value until the excitation is reversed. The total stored energy in the flux bars is small because of the small cross sectional area, even though the field in the bars is high. Flux biasing schemes such as have been used in the betatron are also possible in the betatron phase of the synchrotron, but are not required to reach the 2 Mev transition energy and so are not used.

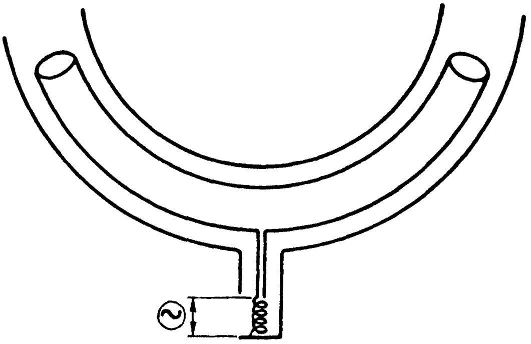
It is possible, of course, to use synchrotron acceleration from the start, in which case orbit radius can be kept constant by modulating the frequency of the electric field during the early part of the accelerating cycle. This type of synchrotron is being constructed at the University of Michigan.[24] With an injection energy of 500 kilovolts the frequency modulation is about 14%, to be obtained by use of a broad band oscillator-amplifier and a low Q resonant cavity. This design also proposes a "race-track" orbit in which the circular magnet is separated into sectors (four quadrants in the final design) with straight sections of vacuum chamber between sectors. One of the straight sections is used for the resonant cavity, another for injection of the 500 kilovolt ions. Theoretical studies[25] of the effect of such magnetic inhomogeneities on ion oscillations show that harmful resonances can be controlled and stability preserved with a minimum of two gaps, and that it improves with the number of gaps. The net effect is an increase in aperture requirements linearly proportional to the additional circumferential length of ion path.

Electron intensities in the synchrotron are small compared with cyclotron ion intensities; they are also very difficult to measure or compute. The average x-ray output from a probe target is the most reliable measure of intensity, and in the General Electric Co. machine is 40 roentgens/minute (at 1 meter) in the center of the x-ray beam. The efficiency of x-ray production at these high energies is not known with certainty, but estimates indicate average electron beam currents of the order of 2×10^{-4} microamperes. The duty cycle can be observed and consists of a sequence of short pulses at the resonant frequency extending over a time interval of a few microseconds, and repeated 60 times/second. It may be possible to deflect an emergent beam of electrons out of the synchrotron. This has been done with the Illinois betatron by using a magnetic "peeler" to weaken the magnetic field in a short sector near the periphery. Other suggestions involve pulsing of disturbing field coils to set up excessive radial oscillations and "slosh" the electrons out of their stable orbits. One of the most useful features of the synchrotron (or betatron) is the opportunity of timing the ejection of electrons so that ejection energies can be controlled. Such a continuously variable maximum energy is of great value in physical experiments.

The radio-frequency cavity used for acceleration has taken several forms in the different design groups. The most popular is a quarter-wave concentric resonator of oval cross section which is shortened physically by using a solid dielectric such as glass or ceramic, and which forms one section of the vacuum wall. In Figure 17a, a typical quarter-wave cavity resonator is shown, with an external feed connecting capacitatively to the internal conductor. The thin metallic conducting walls are electroplated on the surfaces of the dielectric tube and scribed or etched further to reduce eddy current losses. Another type of accelerator is the "C" electrode (equivalent to the "D" in a cyclotron), illustrated in Figure 17b, with dimensions such that it can be driven at the resonant frequency. The electrons are accelerated on entering and also on leaving the electrode.



QUARTER WAVE RESONATOR



"C" ELECTRODE

Fig. 17.—Synchrotron radio frequency accelerating electrodes: (a) Quarter-wave, dielectric filled, concentric cavity resonator (b) "C"-electrode structure which provides two accelerations/revolution.

The synchrotron is subject to the same ultimate energy limitation as the betatron, due to radiation loss. However, it is possible to add more energy per turn to the electrons with the high frequency accelerating electric field of the synchrotron than by betatron induction, and so the limit will be at somewhat higher energies. If necessary, several accelerating gaps could be used to increase the energy/turn. However, eq. 31 shows an extremely rapid increase in radiation loss proportional to the 4th power of the energy. The maximum energy is reached when the radiation loss/turn equals the maximum practical value of acceleration energy/turn. Increasing the radius by lowering the magnetic field can raise this upper limit slightly but will become excessively expensive. Estimates of the maximum practical energy are of the order of 1000 Mev.

VII. THE SYNCHRO-CYCLOTRON

The energy limitation of the normal, fixed frequency cyclotron can be removed and the ions accelerated indefinitely if the applied frequency is varied to match exactly the ion rotation frequency. It was found that this limitation was due to the relativistic increase of mass with energy and the consequent reduction of ion frequency, causing the ions to go out of resonance with the applied electric field. We have also shown that the phase-stable orbits in a cyclotron will follow a slow change in frequency and acquire the correct energy to preserve resonance. If frequency is varied cyclically, a short bunch of ions will be accelerated to high energy in each frequency sweep, resulting in a sequence of such bunches occurring at the modulation frequency. This reduced effective duty cycle results in a lower average ion output than in the conventional cyclotron, but avoids the energy limitation due to a fixed frequency. The primary problem of design for a frequency modulated cyclotron is that of increasing the duty cycle.

The use of frequency modulation as a remedy for the relativistic limitations of the cyclotron was first suggested by McMillan[16] as one of the possible applications of the principle of phase stable synchronous orbits. It was obvious that this method would result in much higher energies than otherwise possible with the 184-inch magnet at the University of California. This magnet had been assembled and used for experimental purposes in the Manhattan District during the war, but not completed as a cyclotron. The first test of the principle, however, was made on the older 37-inch cyclotron magnet[26] by an ingenious method of simulating the expected relativistic mass change with an exaggerated radial decrease in the magnetic field. An 18% modulation in frequency was required to match ion frequencies in this tapered magnetic field, and very small voltages were applied to the accelerating electrode. The success of this model experiment justified a modification of the larger cyclotron to utilize this principle. On completion the 184-inch machine was immediately successful and is now in operation, producing 200 Mev deuterons or 400 Mev He^{++} ions.[18] The relative simplicity of operation as a synchro-cyclotron has impressed all observers with the practicality of this principle. Currently there are six or eight frequency modulated cyclotrons under construction in this country and several others abroad.

The principal feature of the synchro-cyclotron is the solid core magnet required by the continuously expanding orbit (in the constant magnetic field) as energy increases. The relative dimensions of a cyclotron magnet will change only by small factors for different sizes so that the weight of iron (the cost-determinant) will increase almost with the cube of the radius. This rapid increase of cost with energy is the chief limitation on the practical energy. To illustrate this point Table II shows rough estimates of dimensions and magnet weight for several values of proton energy at an assumed magnetic field of 16 kilogauss at the maximum ion radius.

TABLE II. Cyclotron dimensions at 16 kg.

Proton energy (Mev)	Orbit radius (m)	Pole diameter (ft.)	Magnet weight (tons)
100	0.93	7	300
300	1.68	12	1,500
600	2.54	18	5,000
1000	3.55	24	12,000
2000	5.83	40	54,000

The weight of the iron is offset by the relatively simple design and structure of such a D.C. magnet. Cheap iron can be used in the form of thick plates or forgings, exciting coils are simple, even though heavy, and power requirements are not excessive if a sufficient amount of conductor is used in the coils.

The principle of phase stability operates at optimum effectiveness in the synchro-cyclotron. The theory of oscillations in phase and radius follows closely the theory for the synchrotron, except that in this case the radial oscillations occur about a reference orbit of slowly increasing radius. The same description of motion holds that was used in interpreting Figs. 11 and 12. The energy gain/turn depends on the rate of change of frequency with time. As the ions accelerate, the rotation frequency changes according to eq. 22, and as illustrated in the chart of Fig. 10. To achieve a constant equilibrium phase Φ_e, so that the energy gain/turn is constant, requires a nonlinear time variation of frequency. However, this is not essential and a linear frequency-time relation will also result in stable orbits; in this case the phase Φ_e and the average energy gain/turn will change with time. As long as the peak voltage on the electrode is adequate (2 or 3 times the average value), no special care is needed in shaping the frequency-time curve.

The unusual design problem set by the wide range of frequency modulation has been solved by the use of mechanically rotated condensers to resonate with the inductive electrode structure. At Berkeley this condenser consists of a multitoothed disk rotating in vacuum between stator disks of similar shape, located at the outer end of the inductive electrode supports. Electrically this forms a half-wave resonant circuit with the fixed electrode capacity at one end and the variable condenser at the other end. It is used as the load circuit of a self-excited power oscillator, coupled inductively at a point near the node. The relatively high *Q* circuit and low voltages needed keep the power requirements at a reasonable value. The necessary amplitude of frequency modulation is achieved by close spacing of the teeth in the vacuum condenser.

The rate of frequency modulation determines the magnitude of the radio frequency voltage needed. To illustrate take a numerical example of protons being accelerated to 200 Mev in a magnetic field of, say, 16 kilogauss. The initial frequency of rotation (Fig. 10) is 24 Mc and at 200 Mev it is 20 Mc; the reciprocal frequency or time/revolution is (average) 4.5×10^{-8} seconds. Now consider a modulation frequency of 100 cycles/second and assume that the required frequency sweep from 24 to 20 Mc occurs in a quarter of the cycle. So the total time of acceleration is $\frac{1}{400} = .0025$ second and the number of turns is 55,000. To acquire 200 Mev, the average energy gain/turn would have to be 3600 ev. Under optimum conditions, using the theoretical frequency-time function which gives a uniform energy gain/turn, only a slight excess would be needed to maintain resonance; with two accelerations/turn (entering and leaving the electrode) a peak potential of less than 2000 volts would suffice. However, to allow for deviations from optimum conditions, 2 or 3 times this voltage might be required in practice. A faster modulation frequency would raise the voltage requirements proportionately and conversely a slow cycle would reduce them. In Fig. 18 a time plot of frequency variation is shown to illustrate the acceleration interval, using the numerical illustration above to provide the scale.

The Berkeley 184-inch cyclotron can be used to illustrate many special features in the design. Fig. 19 is a schematic drawing of this machine. The walls of the vacuum chamber are in the form of a square with many removable ports on the sides. Top and bottom plates have circular holes which are sealed to the pole tips so that atmospheric pressure is supported by the magnet frame. The single large electrode is supported by heavy insulators at the vacuum wall; the rotating condenser is in a separate vacuum chamber at the ends of the electrode supports. Ions are produced by a discharge tube source similar to that of a conventional cyclotron, mounted at the center of the chamber and

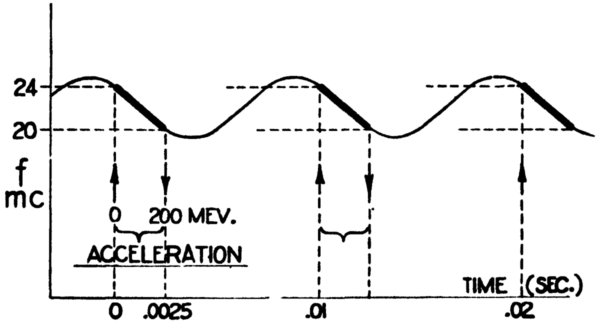


Fig. 18.—Frequency modulation cycle in a synchro-cyclotron. The numbers represent proton acceleration to 200 Mev in a 16 kilogauss magnetic field with a modulation frequency of 100 cycles/second.

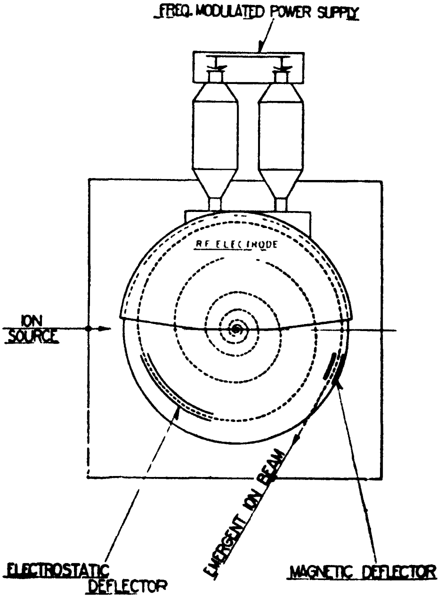


Fig. 19.—Berkeley 184-Inch synchro-cyclotron. Showing single radio frequency electrode with frequency modulating condensers at outer end of the electrode supports, ion source and the electrostatic and magnetic deflectors for the emergent beam.

equipped with vacuum gates for ready replacement of filaments. Probe targets can be inserted from the side to any radial position by means of a sliding vacuum seal. Both electrostatic and magnetic deflecting units are provided to guide the emergent beam. An ion path is shown in the illustration, badly distorted to indicate only a few turns, although in practice the ions make about fifty thousand revolutions.

The magnetic field must be slightly tapered to provide vertical focusing for the ions. This tapering is not extreme; about 5% decrease in 8 feet of radius is sufficient. Nevertheless, this requires an additional 5% of frequency modulation to maintain resonance. Vertical oscillation amplitudes observed at Berkeley are about ±0.5 inch amplitude with such a focusing field. This was observed on probes inserted from the edge and located at various radii. Near the periphery, where the field starts to decrease more rapidly, the vertical oscillations first decrease and then suddenly increase in amplitude and the ion beam impinges against the edges of the electrode and is lost. This "blowing-up" of the beam was found to correspond with the point where the magnetic index $n = 0.2$, for which the radial oscillation is a first harmonic of the vertical oscillation. From eqs. (26a) and (26b) we find:

$$\frac{\omega_r}{\omega_v} = \frac{\sqrt{1 - n}}{\sqrt{n}} \frac{\omega}{\omega} = 2$$

Under these conditions there is sufficient coupling between these two modes of oscillation to allow energy to transfer from the radial mode to the vertical mode. This phenomenon must be avoided in order to pull an emergent beam out of the cyclotron. The radius at which n reaches the value 0.2 is a limiting radius for acceleration and any deflecting system must be located inside this radius. Due to the fringing fields this point is well inside the pole face, at approximately half the gap width in from the edge. This radius and the value of magnetic field at this radius determine the maximum energy available from a magnet. As flux is increased toward the saturation limit the radial location of the $n = 0.2$ position moves inward. So the product BR approaches an asymptotic limit as magnet power is increased. It is not correct to assume a constant maximum radius with increasing field.

To appreciate the problem of deflecting the beam out of the chamber consider again the physical shape of the envelope of resonant ions. This envelope will have a considerable radial width (estimated as 5 to 10 cm. in the Berkeley machine) associated with the radial synchronous oscillations plus free oscillations, an angular spread over a phase angle (possibly $\pi/2$ radians) and a vertical height of 2 to 3 cm. Individual ions will be migrating rapidly inside this envelope due to the free oscillations, and pulsating radially at the much lower phase oscillation frequency. The spacing between successive turns of a single ion is small, too small to use a septum or "beam-splitter" such as is used in the conventional cyclotron.

The technique which has been successful at Berkeley is to use an open deflecting electrode structure consisting of four bars electrically connected in pairs. The inner pair is spaced vertically to allow the ion bunch to pass between them; the outer pair is similarly spaced and parallels the first pair at larger radius. High voltage pulses of short duration (.2 microsecond) are applied, positive on one pair, negative on the other, when the bunch of ions is between the bars, to give a strong radial electric field. The effect of this disturbing field is to set up a large amplitude radial oscillation so that on the next turn the ions are displaced outward by several inches. The electric field is more effective when the force on the ions is inward than the reverse. The sudden expansion of oscillations forces the ions to cross the $n = 0.2$ region rapidly, in a single revolution or less, so that the coupling with the vertical oscillations does not have time to act, and no vertical defocusing occurs. To complete the deflection a short magnetic shield is used, located at the extreme outward location of the shock oscillations. Ions traveling a few feet in the reduced magnetic field of this shield spiral outward rapidly and emerge through the vacuum chamber wall. Approximately 1% of the resonant ion beam can be thus deflected outside the chamber, providing a valuable alternative to the probe targets for many types of experiments.

The Berkeley synchro-cyclotron has produced average ion beam currents of over 1 microampere of 200 Mev deuterons or, alternately, 400 Mev He^{++} ions. On striking a probe target the deuterons split into a proton and a neutron, each of approximately half the energy. Protons of 100 Mev have a radius of curvature half that for deuterons in the magnetic field, and have been studied by placing targets near the center of the cyclotron chamber. Neutrons fly out tangentially from a probe target, since they are uncharged; this beam of 100 Mev neutrons is of exceptional value for research studies.

The average intensity of 1 microampere represents 10^{13} deuterons/second, and on striking a target produces a much larger number of secondary particles and lower energy radiations. The radiation intensity is such that 10-foot thick concrete walls have been installed around the Berkeley machine to provide adequate shielding for the protection of personnel. As an example of the shielding problems, measurements show that 9.5 inches of concrete are required to reduce the intensity of the 100 Mev neutron beam to half value. The shielding problem alone demonstrates the tremendous intensities available from the synchrocyclotron.

Several other laboratories are well advanced on design and construction of large synchro-cyclotrons, including Columbia, Rochester, Harvard, Pittsburgh, Chicago, and several in England, Sweden, and other countries. These are all in the energy range 100 to 350 Mev (for protons). A design study of a 750 Mev machine (20-foot pole diameter) has been undertaken at Brookhaven National Laboratory, which shows the feasibility of design for this energy. Costs rise sharply with energy, however, as indicated in Table II, and it seems probable that 1.0 Bev is about the largest practical size for economic reasons.

VIII. The Linear Accelerator

The linear accelerator is a device for accelerating charged particles to high energies by an oscillating electric field applied to a linear, periodic array of electrodes, with an applied frequency which is in resonance with the motion of the particles. This resonance principle was first proposed by Wideröe[5] and was developed in 1934 by Sloan and Coates[27] and others for the acceleration of heavy positive ions to energies of 1 to 3 Mev. Due to limitations of the radio frequency techniques available at that time, the use was restricted to heavy ions and the utility for nuclear research was insignificant. However, the intensive development of high frequency techniques in the radar field during the past war, particularly the high power, short-pulse magnetrons and the improved understanding and use of wave guides, has increased the potentialities of linear accelerators; several developments are now in progress to explore the possibilities of extending the technique into the hundred Mev range. A thorough analysis of the theory and design of linear accelerators has recently been prepared by Slater[28] as a Technical Report of the Research Laboratory of Electronics at the Massachusetts Institute of Technology. Slater's study gives the theoretical background and describes the progress made since the war in the M.I.T. laboratory; it has been of great value in preparing this summary.

In its earliest form[27] the linear accelerator consisted of a set of tubular metallic electrodes alternately connected to the terminals of a high frequency power supply and enclosed in a glass vacuum chamber. Ions produced in a discharge source at one end were projected along the axis and accelerated on crossing the gaps between electrodes. This tubular structure is illustrated in Fig. 20a. The lengths of the electrodes are arranged to be equal to the distance traversed by the ions in a half-cycle of the electric field, and so increase in length as velocity increases. At low energies (nonrelativistic) the velocity varies with the square root of the energy, so electrode lengths increase in a sequence proportional to the square roots of a series of integers: The electrode length χi for the i^{th} electrode is related to the frequency f, the average accelerating voltage V_e and the e/m value of the ions as:

$$x_i = \frac{v_i}{2f} = \frac{1}{2f}\sqrt{\frac{2eV_e}{m}}\,\sqrt{i} \text{ (nonrelativistic)} \qquad (32)$$

For high energies, where the velocity is essentially equal to the velocity of light, the electrode lengths become constant, and the relation is:

$$x = \frac{c}{2f} = \frac{\lambda}{2} \text{ (relativistic)} \qquad (33)$$

This means that electrodes are spaced at half-wave length intervals and is illustrated by Fig. 20b.

Phase stability is present in the linear accelerator under certain conditions. For particles moving along the axis of the electrode structure with increasing velocity the same type of focusing in phase and energy is obtained as in the synchrotron. There will be one value of voltage across the gaps which is correct for resonance, the value V_e used in eq. 32. In general the peak voltage across the gaps will be greater than V_e and so there will be an equilibrium phase Φ_e for particles to cross the gaps. To show the existence of phase stability consider a particle crossing the gap at another phase illustrated by the point t_1 on the voltage-time graph of Fig. 21, when it finds a higher voltage than V_e. This particle will gain more energy, have a higher velocity and so take a shorter time to reach the next accelerating gap. The resultant phase shifts at the second, third, etc. gaps will reduce the energy/acceleration until eventually the particle will acquire less than the equilibrium value, will fall below equilibrium velocity and the situation will be reversed. This is illustrated in Fig. 21, as the points t_2, t_3, etc. in successive accelerations. Here we have all the requisites for stable oscillation of phase about the equilibrium phase Φ_e.

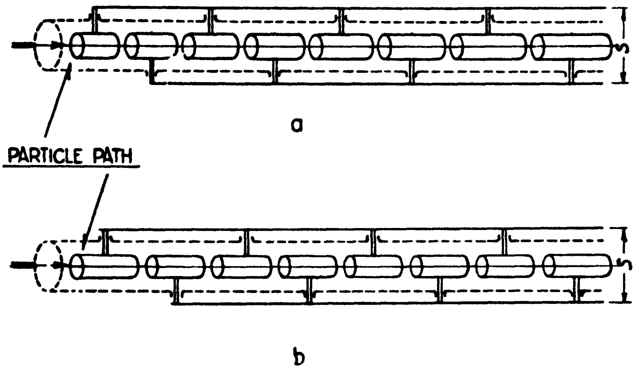


FIG. 20.—Schematic linear accelerator. a) Low energies, where velocity is increasing. b) High energies, where velocity is constant ($=c$).

In the description of phase focusing above, note that the region of stability is centered about a phase where the voltage across the gap is increasing. The alternate position with decreasing voltage would lead to instability. This situation is just the opposite of that for phase stability in the synchrotron, and results in lateral defocusing of the particles. It is one of the most serious handicaps of the linear accelerator that the conditions required for phase stability lead to a lateral spread. To compensate for such a spread, special axial focusing devices are required. Furthermore, even though such lateral focusing is provided the particles thereby traverse paths which deviate from a straight line, travel slightly longer paths and so acquire a phase shift relative to particles which are truly axial.

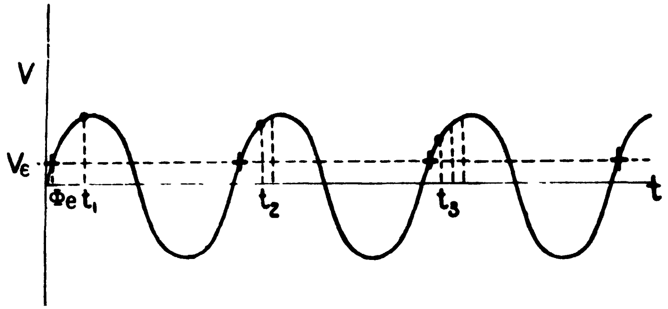


FIG. 21.—Voltage between electrodes in the linear accelerator as a function of time. V_e is the average voltage for resonance; t_1, t_2, t_3, etc. show the phase shifts of a particle in successive accelerations which result in phase focusing.

Electrons are defocused only in the early stages when the velocity is less than the velocity of light and are in neutral equilibrium thereafter. Magnetic lenses around the accelerator tube could probably be used to provide necessary focusing. Protons, however, will be strongly defocused over the entire acceleration. The focusing methods proposed, using foils or grids mounted on the electrodes, have not yet been successfully tested.

The periodic electrode structure can be shown to be electrically equivalent to a wave guide loaded so that the phase velocity of one of the travelling waves is the same as the velocity of the particles. In such a wave guide the electric field will move along the guide so that the particles pick up energy continuously, entirely similar to the motion of a surfboard down the advancing front of a water wave. Post war developments of the linear accelerator in different laboratories have largely been extensions of wave guide techniques and have followed several parallel channels which are closely related theoretically.

For the acceleration of electrons, most laboratories have chosen to capitalize on the developed techniques and high power magnetron sources in the radar field, with frequencies of 3000 Mc. Wave guide loading has been accomplished with iris diaphragms. At the Massachusetts Institute of Technology[28,29] a standing wave type of iris loaded wave guide at 10 cm. wavelength is under development; similar programs are in progress at General Electric Co., the University of Virginia, Purdue University, Yale University, and others. In Figure 22a a schematic representation of the iris loaded wave guide shows the structure and the electric field distribution obtained with standing waves. At Stanford and at the Telecommunications Research Establishment (TRE) in England, the same type of iris loaded wave guide has been used for the propagation of traveling waves. It should be noted that electrons reach a velocity of 0.98 c at 2 Mev energy, and so tapered loading of the guide is needed only at the start; thereafter the sections can be uniformly spaced for the velocity of light.

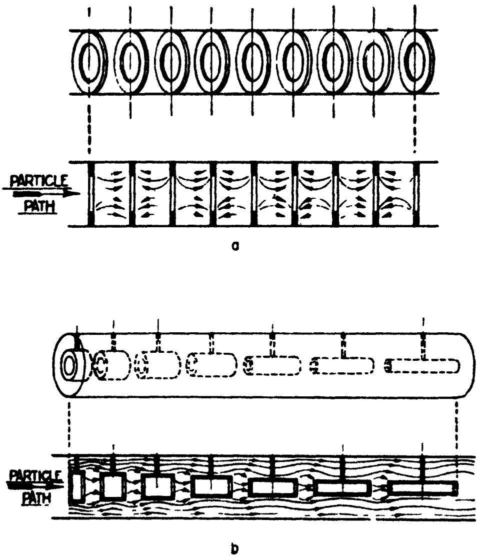


FIG. 22.—a) Iris loaded wave guide for an electron linear accelerator at 3000 Mc showing the geometry and the shape of the lines of electric field intensity. b) Drift-tube wave guide for a proton linear accelerator at 200 Mc showing the geometry and the shape of the lines of electric field intensity.

The wave form propagated in a periodic structure such as is illustrated in Figure 22a can be expressed as a Fourier series, in which only one of the components will have a velocity of propagation equal to the particle velocity. All the other components will have different velocities and will act only as rapidly oscillating perturbing fields so their effects can be neglected. The wave which is propagated with the largest amplitude is that for which the spacing between diaphragms is a half-wave of the guide wave length; this is called the "π-mode." If the structure is terminated after a finite number of diaphragms, so that the waves are reflected from the ends, a standing wave is set up along the guide, consisting of two sinusoidal components which are out of phase by 180°. The guide acts like a resonant cavity for this frequency, with diaphragms located at the nodes of the standing wave pattern. Under these conditions the velocity of propagation can be either equal to or smaller than c, and can be adjusted by the spacings of the diaphragms and the size of the apertures. The axial component of the electric field, which must be integrated along the guide to give particle energies, is proportional to the stored energy and so to the power input. Losses in the guide produce attenuation and reduction of the stored energy, and are additive for all the possible modes (this subject is discussed in some detail by Slater) Using the high peak pulse power possible with commercially available magnetrons, it is possible to achieve fields capable of accelerating electrons at the rate of 1 to 2 Mev/foot of guide for short lengths. The average potential drop/unit length $E\chi$ is a measure of the quality of the design.

Another way of using the loaded wave guide is by employing a non-reflecting termination and obtaining a traveling wave which can be considered as consisting of a sine wave and a cosine wave with a phase difference of 90°. These must be excited separately, by feeding power into at least two points, and phased in quadrature. The time required for build up of the field is finite, and so a short accelerator fed at one end will not give the maximum amplitudes possible with a longer one. On the other hand, if the guide length is long enough to give a transit time long compared with this build-up time, one end does not know what the other end is doing, fields can build up to maximum and power can be fed in at uniform intervals along the guide. The traveling wave accelerator has the advantage of utilizing a larger fraction of the power available from a magnetron oscillator than a standing wave tube. However, losses in the guide are greater due to the large number of diaphragms and the net acceleration is not significantly better than for standing waves at the same input power.

Several factors influence the acceleration/unit length, given by the field $E\chi$:

(a) $E\chi \simeq P\chi^{\frac{1}{2}}$, where $P\chi$ is power input/unit length.

(b) $E\chi \simeq \lambda^{-\frac{1}{4}}$, where λ is wavelength. This is primarily the effect of skin depth on the G_c of the guide, and shows a slow improvement with decreasing wavelength or increasing frequency.

(c) Geometry of apertures in iris diaphragms: The effect of small apertures is to concentrate the stored energy in the region near the axis, thereby increasing the field on the axis. However, a finite aperture is required for the particle beam and so there must be a compromise choice for each accelerator design.

(d) Operating mode: The π-mode, illustrated in Fig. 22, has diaphragms set at $\lambda/2$ intervals, has a high G_0, is ideal for standing waves, but cannot support a traveling wave. The $\pi/2$-mode has twice as many diaphragms, set at 90° phase intervals and so has greater surface losses and lower G_0; however, it does allow a traveling wave and has other compensating advantages.

Since electrons acquire a velocity essentially equal to the velocity of light in the first few feet of acceleration there will be relatively weak phase focusing. The electrons will stay in resonance in the standing wave type of accelerator or in the accelerating phase of the traveling wave type only if the phase velocity in the guide is exactly the velocity of light. This depends on the precision of construction of the guide and the precision of frequency and phasing of the power sources. The frequency of separate self-excited oscillators is hard to regulate with precision; power amplifiers with a master oscillator to determine frequency would be more amenable to control, at least for the traveling wave tube. The precision required is directly related to the length of tube and so to the maximum energy. At 10 cm. wavelength and with 1 Mev per foot acceleration, there will be roughly 300 wavelengths for 100 Mev and the frequency should be controlled to better than 1 part in 10^4. Bunching of electrons by phase focusing at low energies will be preserved but not improved at relativistic energies. Electrons will be accelerated over a considerable range of phase and so have a considerable final energy spread. A good estimate of this energy spread is that half of those accelerated will have energies within 10% of the maximum energy.

The linear accelerator for protons designed and built by Alvarez[30] at the University of California uses a frequency of 200 Mc, originally chosen because of the availability of surplus radar equipment at that frequency. The structure is a wave guide with "drift tubes" of the proper resonant lengths mounted along the axis. A diagram of the structure and a sketch of the lines of electric field intensity is shown in Figure 22b. The drift tubes increase in length and decrease in diameter to produce a phase velocity equal to proton velocity during acceleration. Any unit section of the guide can be viewed as a cavity resonant at 200 Mc and having a distance between accelerating gaps compatible with the proton transit time. The copper wave guide is about 4 feet in diameter and 40-feet long, enclosed in a horizontal cylindrical steel vacuum chamber. Protons from a 4 Mev electrostatic generator are injected at one end and travel down the axis, acquiring an additional 28 Mev energy from the 30 gaps between drift tubes. Defocusing of the ions by the electric fields in the gaps was to be eliminated by using thin metallic foils over the entering faces of the drift tube apertures. The field between a cylinder and a plane is so shaped as to be focusing for particles penetrating the plane. Foils must be so thin that the energy loss is small compared with the energy acquired in the acceleration. Reports of preliminary tests show that the ¼-mil beryllium foils used were burned out by high frequency discharges, and so grid structures are being developed to provide the necessary focusing.

In considering the extrapolation of linear accelerators to the 100 to 1000 Mev range the most difficult problem is that of feeding power to the wave guide. This is essentially the problem of operating many parallel power sources at the same frequency and the same phase. Using short lengths of guide several magnetron oscillators have been phased together successfully, arranged so that the power is distributed uniformly along the guide Extension of this technique to many oscillators is not simple, and is one of the primary design problems For the standing-wave accelerators this power will be dissipated uniformly in the walls of the guide and sets a constant figure for the power/unit length required. With traveling waves there is also a flux of power out the end of the guide, decreasing the energy/unit length available to the particles. This factor becomes of less importance for long lengths, where, however, the problems of phasing become more difficult. It is too early to say what success may be achieved by use of new techniques now under development. So the use of linear accelerators in the supervoltage range must wait for the solution of the problems of multiple power sources.

IX. FUTURE POSSIBILITIES: THE PROTON SYNCHROTRON

The machines described in the preceding sections have all been proven practical by actual construction and operation. They are limited in their potentialities, however, and none of them offer much hope for extension above about 1000 Mev (1 Bev or billion electron volts). Scientific research in the future may well require particle energies in excess of this limit. One reason is that the probability of production of mesons in the 300 Mev range of existing machines may prove to be so small that results may be inconclusive. It may also be necessary to produce mesons in pairs or in showers (observed with high energy cosmic ray primaries) in order to understand completely the production processes. Eventually it may be necessary to exceed the energy associated with the rest mass of a proton or neutron (about 1 Bev) in order to study the details of nuclear forces. To produce a pair of neutrons, 2 to 6 Bev of kinetic energy would be required, depending upon the choice of bombarding particle. So some studies have been directed towards the design of machines for greater than 1 Bev.

Linear accelerators are in too early a stage of development to predict their possibilities in the Bev range; at 1 Mev/foot, a 1 Bev accelerator would be 1000-feet long, an obviously difficult problem of construction, alignment, and phasing. Electrons will reach the radiation loss limit in magnetic accelerators at about 1 Bev. So the only available method of producing more than 1 Bev with present knowledge is by using protons and circular orbits in a magnetic field. Due to excessive weights and costs of the solid core cyclotron magnet, a ring-shaped magnet seems essential. The resulting machine would be a proton synchrotron, utilizing both variable magnetic field and variable frequency. Construction of a 1.3 Bev proton synchrotron is in progress at the University of Birmingham, England,[81] and a theoretical analysis of ion orbits and design requirements has been published.[82] Design studies for even higher energy machines of this type are in progress at Brookhaven National Laboratory and the University of California.

In a proton synchrotron the magnetic field applied to the ring-shaped orbit would be modulated from low to high field intensities, probably in a cyclic manner. Power requirements will limit the rate of rise of the magnetic field, so that the acceleration time may be of the order of 1 second duration. The magnet might be cycled at the rate of 1 to 10 cycles/minute. Protons would be injected at low energy, when the magnetic field is correct for motion in the circular orbit at injection energy. Energy would have to be added to the protons at the rate prescribed by the time rate of change of magnetic field. Frequency of ion rotation varies with the increasing velocity of the ions, so the frequency of the accelerating electric field must be modulated to correspond exactly with this resonance frequency. The range of frequency modulation is determined by the velocity at injection and at maximum energy; in the Birmingham design this is 0.27 Mc (300 kilovolts) to 9.5 Mc (1.3 Bev), a frequency ratio of 30:1. Phase focusing applies exactly as in the synchrotron if the applied frequency is correct, and the same stable oscillations occur in phase, energy and radius of path. However, in this case some additional technique must be developed to control frequency and maintain the correct value. This special requirement will probably prove to be the most critical feature in the design of a proton synchrotron.

The engineering design will involve several problems of large magnitude. A compact and efficient magnet design is essential, to reduce cost in this most expensive component, for which thousands of tons of laminated iron are needed. Techniques must be developed for storing and cycling the extremely large amount of stored energy in the magnetic field. The best guess at present is the use of rotating machinery and large flywheels such as have been developed for steel rolling mills. Electrical controls of the high currents to switch the energy from flywheels to magnet and back are a large part of this problem. Another problem new to the engineering field is the production and control of high radio frequency power over a frequency range of greater than 10 to 1. Resonant circuits can be tuned over this range only with great difficulty. On the other hand the power requirements for an untuned electrode circuit are excessive. New concepts and techniques must be developed to solve these problems.

Ion injection should be at energies as high as practical in order to reduce the range of frequency modulation. Present designers visualize the use of an electrostatic generator operating at about 4 Mev as a possible source. It would also be preferable to have the protons injected in preformed bunches, synchronized with the accelerating electric field. Ions can be injected into the orbit by means of suitable deflecting electrodes, which might also be pulsed to minimize interference with rotating ions. The number of ions which can be accepted from such a source in the phase acceptance intervals of the synchrotron will be small, compared with cyclotron intensities, but probably more than adequate for most of the experimental purposes visualized. There is also the danger of serious reduction of intensity by gas scattering during the transit of the millions of revolutions and the hundreds of thousands of miles traversed during acceleration. This puts stringent requirements on the design of the doughnut-shaped vacuum chamber and the pumping system. It is too early to predict the success of ejection systems, but these will probably be based on the pulsed deflector used in the synchro-cyclotron. The hope would be to obtain a narrow beam of monokinetic protons, deflected tangentially out of the orbit, and timed to give any desired energy up to the maximum.

Basic dimensions and ion frequencies can be obtained from the charts of Figs. 9 and 10. To illustrate the sizes involved these basic dimensions are given in Table III for two sizes which have been considered in design studies:

TABLE III. Proton synchrotron.

Max. En (15 kg.)	2 7 Bev	10.0 Bev
Orbit radius	25 ft.	80 ft.
Inject. frequency (4 Mev)	0.58 Mc	0.18 Mc
Max. frequency	6.4 Mc	2.0 Mc

A qualitative sketch of a 25-foot proton synchrotron is shown in Figure 23.

The physical principles of the proton synchrotron appear to be entirely sound, and the engineering problems are soluble even though difficult. No other method that has been visualized up to the present offers the same potentialities. If the proper effort is put into development, a multibillion electron volt accelerator could probably be built within a few years. The scientific fruits of such a machine would certainly justify the effort, and might be of outstanding importance.

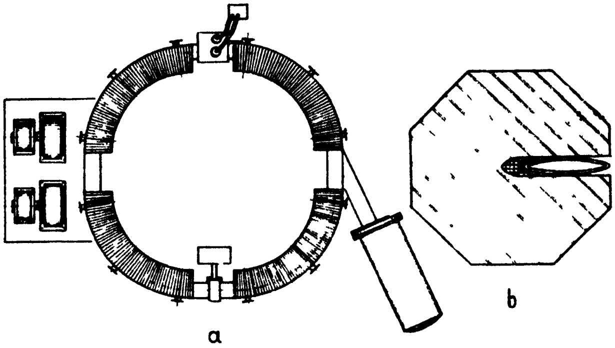


FIG. 23.—Proton synchrotron design. a) Plan of "race-track" magnet with electrostatic generator ion source, accelerating system and vacuum pumps. b) Enlarged section of magnet and vacuum chamber.

REFERENCES

1. Cockcroft, J. D., and Walton, E. T. S. *Proc. Roy. Soc.*, **136A**, 619 (1932); **137A**, 229 (1932).
2. Van de Graaff, R. J. *Phys. Rev.*, **38**, 1919 (1931).
3. Buechner, W. W., Van de Graaff, R. J., Sperduto, A., McIntosh, L. R., and Burrill, E. A. *Rev. Sci. Instrum.*, **18**, 754 (1947).
4. Herb, R. G., Turner, C. M., Hudson, C. M., and Warren, R. E. *Phys. Rev.*, **58**, 579 (1940).
5. Wideröe, R. *Arch. Elektrotech* , **21**, 387 (1928).
6. Lawrence, E. O., and Edlefsen, N. E. *Science*, **72**, 376 (1930).
7. Lawrence, E. O., and Livingston, M. S. *Phys. Rev.*, **40**, 19 (1932).
8. Livingston, M. S. *J. Appl. Phys.*, **15**, 2, 128 (1944).
9. Livingston, M. S. *Rev. Mod. Phys.*, **18**, 293 (1946).
10. Lawrence, E. O., Alvarez, L. W., Brobeck, W. M., Cooksey, D., Corson, D. R., McMillan, E. M., Salisbury, W. W., and Thornton, R. L. *Phys. Rev.*, **56**, 124 (1939).
11. Kerst, D. W. *Phys. Rev.*, **58**, 841 (1940).
12. Westendorp, W. F., and Charlton, E. E. *J. Appl. Phys.*, **16**, 581 (1945).
13. Kerst, D. W. *Phys. Rev.*, **60**, 47 (1941).
14. Kerst, D. W., and Serber, R. *Phys. Rev.*, **60**, 53 (1941).

15. Blewett, J. P. *Phys. Rev.*, **69**, 87 (1946).
16. McMillan, E. M. *Phys. Rev.*, **68**, 143 (1945).
17. Veksler, V. *J. Phys. USSR*, **9**, 153 (1945).
18. Brobeck, W. M., Lawrence, E. O., MacKenzie, K. R., McMillan, E. M., Serber, R., Sewell, D. C., Simpson, K. M., and Thornton, R. L. *Phys. Rev.*, **71**, 449 (1947).
19. Goward, F. K., and Barnes, D. E. *Nature, Lond.*, **158**, 413 (1946).
20. Elder, F. R., Gurewitsch, A. M., Langmuir, R. V., and Pollack, H. C. *J. Appl. Phys.*, **18**, 810 (1947).
21. Dennison, D. M., and Berlin, T. H. *Phys. Rev.*, **70**, 58 (1946).
22. Bohm, D., and Foldy, L. *Phys. Rev.*, **70**, 249 (1946).
23. Frank, N. H. *Phys. Rev.*, **70**, 177 (1946).
24. Crane, H. R. *Phys. Rev.*, **69**, 542 (1946).
25. Dennison, D. M., and Berlin, T. *Phys. Rev.*, **70**, 764 (1946).
26. Richardson, J. R., MacKenzie, K. R., Lofaren, E. J., and Wright, B. T. *Phys. Rev.*, **69**, 669 (1946).
27. Sloan, D. H., and Coates, W. M. *Phys. Rev.*, **46**, 539 (1934).
28. Slater, J. C. Tech. Report No. 47, Res. Lab. Electronics M. I. T., Sept. 2, 1947.
29. Halpern, J., Everhart, E., Rapuano, R. A., and Slater, J. C. *Phys. Rev.*, **69**, 688 (1946).
30. Alvarez, L. W. *Phys. Rev.*, **70**, 799 (1946).
31. Oliphant, M. L., Gooden, F. S., and Hide, G. S. *Proc. Phys. Soc. Lond.*, **59**, 666 (1947).
32. Gooden, J. S., Jensen, H. H., and Symonds, J. L. *Proc. Phys. Soc. Lond.*, **59**, 677 (1947).

Ionospheric Research

A. G. McNISH

Central Radio Propagation Laboratory, National Bureau of Standards, Washington, D. C

CONTENTS

		Page
I.	Introduction	317
II.	Research During World War II	320
III	Geomagnetic Effects in the F2 Layer	321
IV.	Distribution of E and F1 Layers	324
V.	Two Control-Point Method of Calculating Maximum Usable Frequencies	324
VI.	Effects of Solar Activity	326
VII.	Prediction of Ionospheric Disturbances	330
VIII.	Sporadic E Reflections	332
IX.	Absorption of Radio Waves	333
X	Radio Noise	338
XI.	Reflections from Meteor Trails	340
XII	High-Speed Multifrequency Recorder.	342
XIII.	Trends of Research	343
	References	343

I. Introduction

First apprehension of the existence of an electrically conducting region in the atmosphere is due to Balfour Stewart who in 1882 hypothesized the existence of such a region to account for the diurnal variations in geomagnetism. Schuster fashioned this idea into a formal theory in 1889, attributed the electric conductivity to action of the sun's ultraviolet light, and proposed a law for the variation of conductivity which resembles in many features present theories of the formation of the ionosphere.

After Marconi succeeded in spanning the Atlantic by radio in 1900, both Kennelly and Heaviside, apparently unaware of the earlier work of Schuster, proposed independently in 1902 that the bending of the radio waves around the earth is due to the presence of a reflecting layer in the atmosphere. The next two decades witnessed extensive theoretical development in the field of radio wave propagation. Some investigators approached the problem entirely from the classical viewpoint; others, drawing heavily on the work of Lorentz, derived the theory of the propagation of electromagnetic waves in ionized media. Meanwhile the practical aspects of radio wave propagation remained largely empirical. During all this time there had been no direct experiments to determine the nature of the ionosphere.

Experimental ionospheric research began with the experiments conducted independently in 1925 by Appleton and Barnett in England and by Breit and Tuve in the United States. The Appleton-Barnett experiment consisted of the transmission of a continuous frequency and the measurement of the interference pattern set up by the direct (ground) waves and the waves reflected by the ionosphere. The method of the Breit-Tuve experiment involved the transmission of a pulse of radio waves and the measurement of the time delay between the emission of the pulse and the reception of its reflection from the ionosphere. Although the cw technique is still used for some special experiments the pulse technique has come to be the standard method for present day probing of the ionosphere.

The pulse technique furnishes very concise information on the ionosphere. The frequency of a wave reflected from a certain height is a measure of the ion (or electron) density at the height at which reflection occurs, the frequency reflected being proportional to the square root of the ion density. If the frequency exceeds a certain critical value it will not be reflected. This critical frequency is a measure of the maximum ion density for a region. The time delay between emission of the pulse and reception of the reflection is a direct measure of the virtual height of the reflecting region from which the true height may be calculated by correcting for retardation of the wave packet in the ionized region. The critical frequency and virtual height are fundamental parameters for analysis of a radio communication problem.

Within a few years regular observations of the ionosphere using the pulse technique were initiated at several different stations. At first these observations consisted of the measurement of the time delay of reflections returned from the ionosphere on one or two fixed radio frequencies; later measurements were made throughout the spectrum of reflectable frequencies by manual step-wise variation of the frequency. Automatic methods of continuously varying the frequency and photographically recording the reflections were developed at the National Bureau of Standards, incorporated into practice in 1933, and placed in regular operation at the Bureau's field station near Washington. Similar equipment was also developed later at the Department of Terrestrial Magnetism, Carnegie Institution of Washington, and installed at the Institution's magnetic observatories at Watheroo, Australia, and Huancayo, Peru. Programs of ionospheric research were initiated in other parts of the world, but the observations at the above-mentioned locations probably present the most significant data obtained during this stage of ionospheric research.

From these and other observations made during the 1930's many of the most significant characteristics of the ionosphere were revealed or suggested. Instead of consisting of a single, relatively thin layer of ionization, as was supposed by the earlier investigators, it was clear that the ionosphere is an extensive region beginning at a height of something less than 100 km. and extending many hundreds of kilometers out into space. These observations showed that the ion density increases with height, but not monotonically. During daylight hours there are ordinarily three heights at which the gradient of ion density becomes zero. The designations, E layer, F1 layer, and F2 layer, were assigned to these regions which have their heights of maximum ion density at slightly over 100 km., about 200 km., and about 300 km., respectively. A region below the E layer, thought to be responsible for radio wave absorption, was designated the D region.

It was recognized that the diurnal variation and geographical distribution of the E and F1 layer are such that they can be fairly well represented by a simple function of the sun's zenith angle and that these layers undoubtedly are due to photoionization by the sun's ultraviolet light. The diurnal variation of the F2 layer, on the other hand, does not exhibit close dependence on the sun's zenith angle. At Huancayo, for example, the F2 ion density is less around noon than it is during the midmorning and midafternoon hours. When the series of data available for Washington was supplemented by the series of data from Watheroo, which is located in a corresponding southern latitude, it became evident that the geographical distribution of the F2 layer is not what would be expected from a simple photo-ionic process. The Washington data had shown a winter maximum of ion density instead of a summer maximum which was temporarily explained as a result of thermal expansion of the atmosphere, but the Watheroo data showed maxima at the equinoxes which was inconsistent with such an explanation.

Most of the erratic characteristics of the ionosphere were revealed by these preliminary observations. Among these are the occurrence during both night and day of sporadic reflections from the E region at frequencies above the normal critical frequency of that region. Disturbances of the ionosphere associated with geomagnetic storms had also been observed and their outstanding characteristics, depression of the F2 critical frequencies and development of low-lying absorption layers, had been clearly defined. Sudden ionospheric disturbances involving fade-out of reflected radio signals on the daylight side of the earth had been associated with their cause, bright eruptions in the solar chromosphere. Increases in ionization throughout the ionosphere with increases in solar activity had been discovered and considerable study had been devoted to the effects of solar eclipses on the ionosphere. Many other effects of lesser importance at the time had been noted. All of this fund of knowledge concerning the ionosphere available at the time has been covered in several reviews which appeared just prior to the war.

But application of this scientific knowledge to the practical problems of radio communication lagged. Although the theory for applying the scientific facts to the problems of communication was fairly well understood, lack of knowledge of the world-wide distribution of the F2 layer, which is so important in long-range communication problems, prevented description of the characteristics of propagation paths over most of the earth. Communication engineers relied more upon past performance of radio circuits than upon inferences drawn from the extremely limited scientific information.

II. Research During World War II

Since reliable communication is a primary requirement of military operations, and since radio was the basic method of communication during the far-flung campaigns of World War II, both the Allied and the Axis powers embarked on extensive programs of ionospheric research after the outbreak of hostilities. Central laboratories were established in the various countries and extensive networks of field stations (see Fig. 1) were set into operation to acquire data on the ionosphere in places for which there was none. The purposes of these organizations were to determine the characteristics of the ionosphere in various parts of the world and from them forecast the normal diurnal and seasonal variations, and the long term trends following the sunspot cycle. Improved methods for interpreting the data were developed and more exact relationships were derived for applying the data to forecasting radio propagation conditions. Programs of special solar observations were initiated in order to predict as accurately as possible the occurrence of disturbances of the ionosphere which were known to be of solar origin.

This program accomplished its primary purpose. Great improvement was achieved in the prediction of propagation conditions even for regions in which there was no extensive previous communication experience. The previously irregular distribution of F2 ionization was found to fit into a regular system so that its characteristics could be predicted where they had not been observed. But there was little advance along purely theoretical lines. The tremendous quantity of data collected, the analysis and systemmatizing of these data, and their practical applications to the problems of radio communication constitute the advances in ionospheric research during the wartime period.

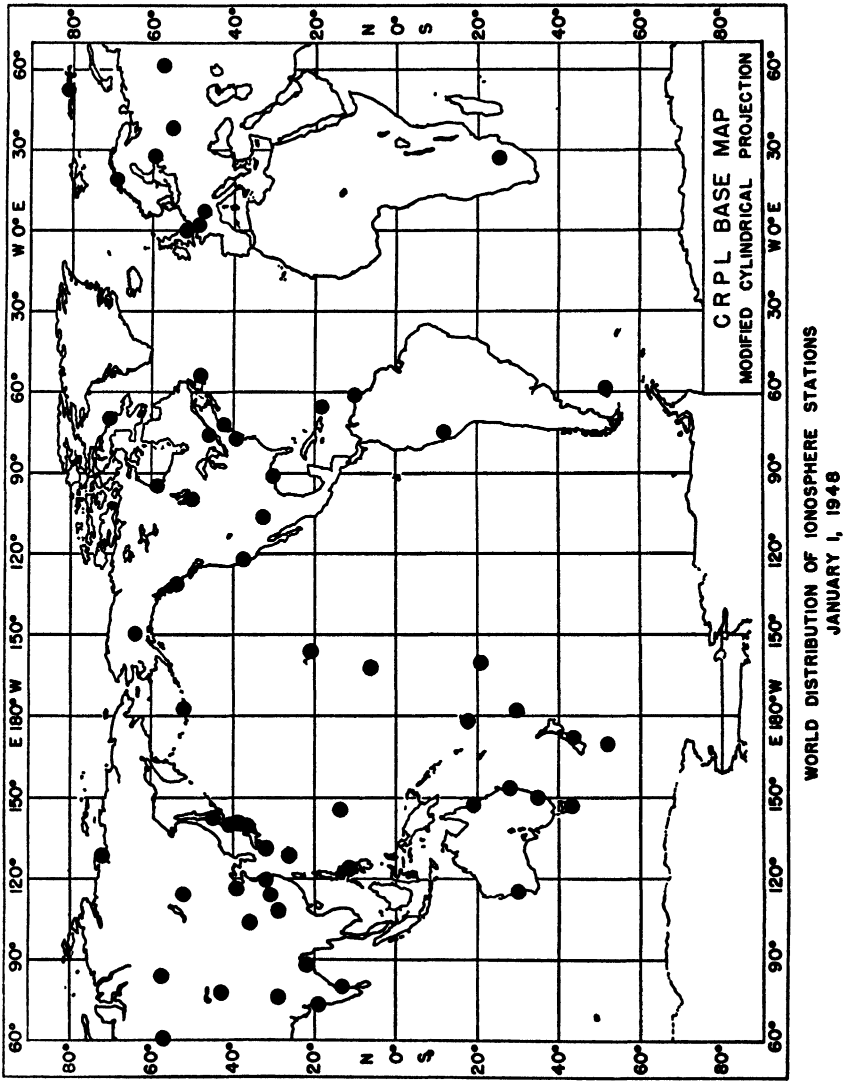


FIG. 1.—Locations of operating ionospheric stations as of January 1, 1948.

III. GEOMAGNETIC EFFECTS IN THE F2 LAYER

One of the most outstanding results of the world-wide survey of ionospheric characteristics was the revelation of a close correlation between the distribution of F2 ion density and the geomagnetic field. This discovery immediately systematized in part the apparently anomalous behavior of the F2 layer. Behavior of the F2 layer at stations in the same geomagnetic latitude exhibits more similar characteristics than at stations situated in the same geographical latitudes. (Geomagnetic latitude refers to a system of spherical coordinates based on the axis of uniform magnetization of the earth which intercepts the surface in latitude of 78.5° N, longitude 69° W.)

For this reason charts of critical frequency vs. local time are not applicable for all longitudes. Fairly satisfactory representations have been obtained, as a temporary expedient, by dividing the world into three zones in the form of lunes with their apexes at the poles of the geomagnetic axis, the west zone covering approximately the region 60° east and west of the 69° W meridian, the east zone covering approximately the region 60° east and west of the 111° E meridian, and the double intermediate zone covering the remainder. Use of separate F2-layer charts for different longitudes for radio propagation purposes was introduced by the Interservice (now Central) Radio Propagation Laboratory of the National Bureau of Standards in 1943. (See Fig. 2.)

A distinct feature of this geomagnetic effect is a lower value of ion density in the F2 layer around local noon in the region close to the geomagnetic equator as compared with values further north and south. A simple photo-ionic control of the F2 ion density would call for highest values near the geographic equator without any special variation associated with the geomagnetic equator.

An interesting aspect of this geomagnetic effect is revealed in correlations of daily fluctuation in the F2 critical frequencies (which are proportional to the maximum ion density) at different stations. Appreciable departures from the running mean value occur in the daily values of F2 critical frequency at a given hour (standard deviation equals 10%, approximately). The correlation coefficient of these departures for pairs of stations in the same longitude but in different latitudes is high for pairs of stations in middle latitudes and decreases as the distance between stations increases. However, the correlation coefficient relating these departures for pairs of stations, one of which lies close to the geomagnetic equator and the other of which lies some 20° to the north or south, is negative.

The implication of this surprising relationship is that a portion of the departure from the running mean value is due to an agency affecting the earth as a whole, such as a change in intensity of the solar ionizing agent, and that another portion of the departure is of terrestrial origin which acts differently in different latitudes. It has been suggested that this terrestrial agent is the diurnal variation in the geomagnetic field which generates electric forces that act on the electric charges in the ionosphere,

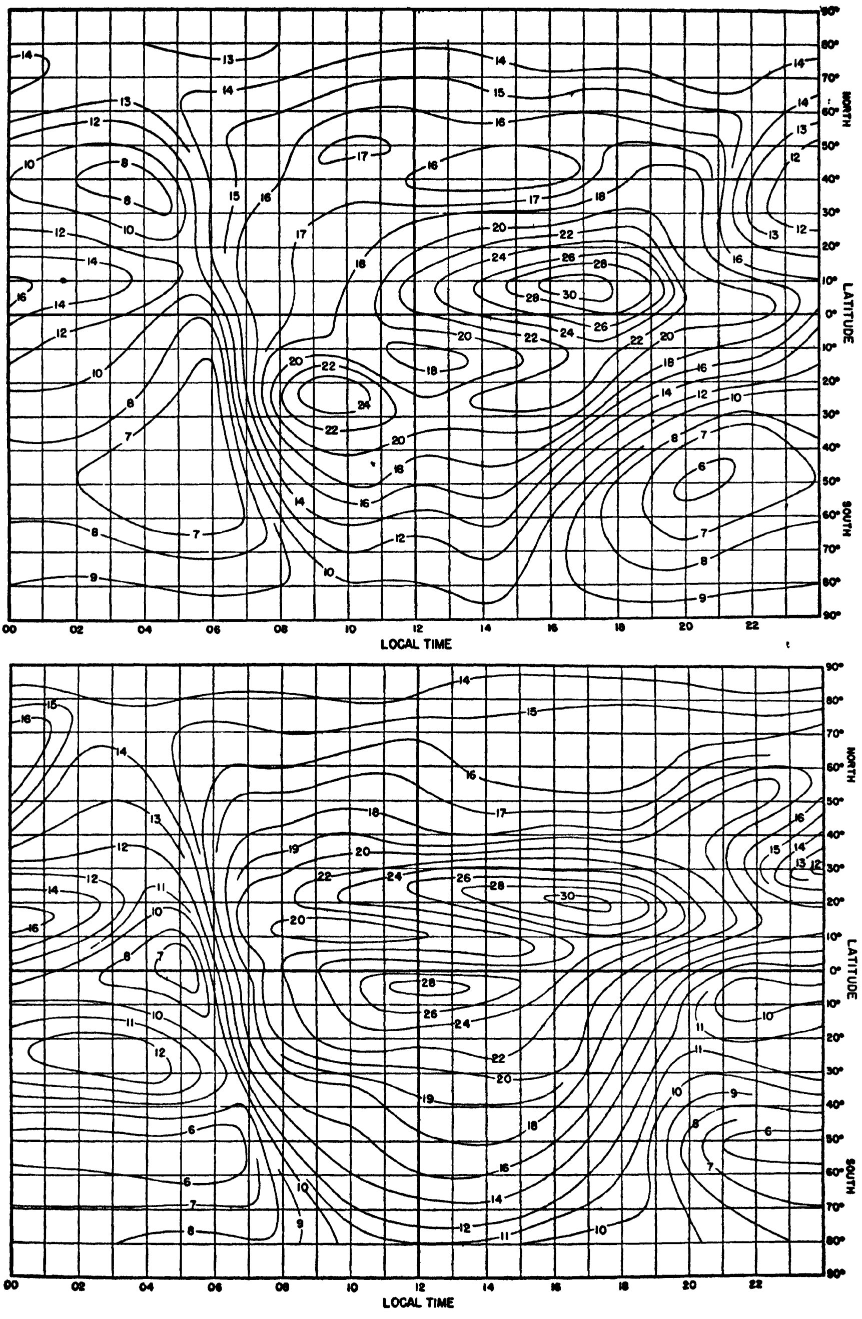


FIG. 2.—Maximum usable frequency charts showing maximum frequencies which will be propagated by the ionosphere for distances in excess of 4000 km. as functions of latitude and local time at point of reflection during June, sunspot number zero, west zone (upper chart) and east zone (lower chart).

but the underlying theory is highly complicated and has not yet been quantitatively examined.

IV. DISTRIBUTION OF E AND F1 LAYERS

The abundant data supplied by the world-wide network of ionosphere stations established during the war permits a more precise determination of the relationship between ion density and the sun's zenith angle, χ. This, in turn, furnishes a basis for inference regarding the processes of ionization and ionic decay in the layers, particularly in the E and F1 layers.

According to conventional theory of layer formation, which assumes that the recombination coefficient is invariant with height, the maximum ion (electron) density should be proportional to $\cos^{\frac{1}{2}} \chi$, and consequently the critical frequency proportional to $\cos^{\frac{1}{4}} \chi$, if the processes are sufficiently rapid that a state of approximate equilibrium exists. On the other hand, if the recombination coefficient is a function of height, linearly dependent on the pressure, or, if the removal of electrons, which are mainly responsible for reflections, is accomplished by attachment to heavy neutral molecules, the electron density will be proportional to $\cos \chi$ and the critical frequency to $\cos^{\frac{1}{2}} \chi$. Such a layer, of course, cannot have a maximum of electron density, unless the constituent gases of the atmosphere have certain complicated distributions, so that the *a priori* likelihood that the ionic decay processes in the E and F layer behave in this manner is low. On the other hand, since no maximum ion density appears in the D region, it may be that attachment is an important factor there. This is consistent with the existence of higher molecular densities in the D region which is more favorable to attachment processes.

Examination of the world-wide distribution of the F1 critical frequency shows that it does not depart conspicuously from the $\cos^{\frac{1}{4}} \chi$ law. E layer critical frequency is approximately proportional to $\cos^{\frac{1}{3}} \chi$. This suggests that either the recombination coefficient varies with height but less rapidly than the pressure, or that both the attachment and the recombination processes are important in the ionic decay processes of the E layer.

V. TWO CONTROL-POINT METHOD OF CALCULATING MAXIMUM USABLE FREQUENCIES

Application of ionospheric data to solution of propagation problems has usually been performed on the basis of a ray theory. A radio wave propagated vertically upward is reflected at a height where the ion density has a value N given by the equation

$$N = \pi m f^2 / e^2,$$

where f is the frequency of the radio wave, m, the mass in grams of the ions (usually free electrons) responsible for the reflection, and e, the electronic change in electromagnetic units. Actually, the presence of the earth's magnetic field causes splitting of the wave into two components, the ordinary, f_o, and the extraordinary, f_x, rays. The above relation applies for the ordinary ray. The frequency for the extraordinary ray reflected at the same ion density is given by

$$f_x^2 - f_x f_H - f_o^2 = 0,$$

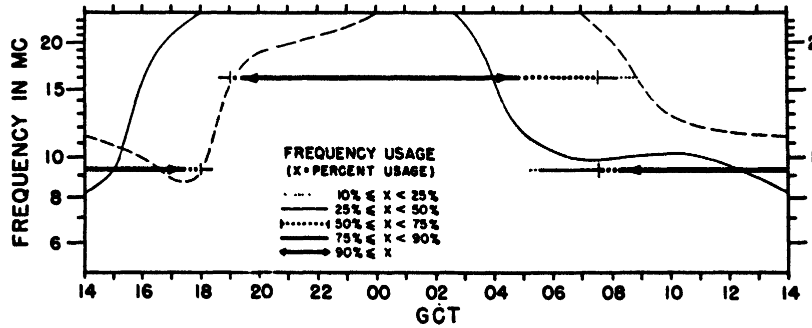
or

$$f_x = f_o + f_H/2$$

for values of f_x and f_o large as compared to f_H, the gyromagnetic frequency of the ions. At oblique incidence on the ionospheric layers a wave is reflected from the same level of ionization if its frequency is

$$f' = f \sec \varphi$$

where φ is the angle of incidence on the layers. This frequency f' is the maximum usable frequency for communication between points where φ



COMPARISON OF PREDICTED FREQUENCIES WITH FREQUENCIES ACTUALLY USED SYDNEY–SAN FRANCISCO OCTOBER 1944.

FIG. 3.—Comparison of predicted usable frequencies with those actually used between San Francisco and Sidney. The solid curve shows frequency limited by San Francisco control point, and the dotted curve, Sidney control point.

satisfies the geometrical considerations involved, applying for the ordinary or extraordinary ray accordingly as f is the ordinary or extraordinary ray. For radiation emitted at very low angles, propagation to a distance of about 4000 km. may be accomplished by a single reflection from the ionosphere in accordance with the ray theory; for greater distances two or more reflections from the ionosphere must occur.

Over long paths many modes of propagation may be involved. Estimating the frequency to be used for communication over such a long path would require calculation of the usable frequency at each ionospheric reflecting point for each of the various possible modes. However, it has been found that propagation of the wave between two points occurs if the conditions for reflection, as given in the above equations, are satisfied at two places along the great circle path between the points, one approximately 2000 km. from the point of emission and the other 2000 km. from the point of reception. These are known as the control points of the path. (See Fig. 3.)

No satisfactory theory of the two-control point principle has been offered. Application of the method to the analysis of propagation paths is based on its empirical success. Success of the method casts some doubt on the adequacy of the ray model for explanation of long distance propagation and may call for formulation of a wave guide model to replace it.

VI. Effects of Solar Activity

The relation between solar activity and ion density of the various ionospheric layers is more completely represented now than hitherto. This is the result of the present availability of more extended series of data, including, in particular, the data from the present sunspot cycle during which the relative sunspot number has reached higher values than have ever been attained since regular observations of the sunspots were instituted. Variations of critical frequencies of the various layers of the ionosphere in relation to the sunspot cycle have been studied at a large number of stations. Particular attention has been devoted to the irregular characteristics of the F2 layer. Its behavior at a large number of stations as a function of various variables has been extensively studied.

Apart from their scientific significance, these studies have been necessary for forecasting propagation conditions in various parts of the world. If the trend of critical frequencies as a function of solar activity is determined, then an estimate of future solar activity fixes an estimate of the maximum usable frequency which may be employed for radio traffic between any two points. This, of course, implies the ability to estimate future trends of solar activity which may be achieved, with some degree of success, on a statistical basis from the lengthy series of sunspot observations available today.

The presently available series of critical frequency and sunspot observations may be fairly well related by an equation of the form

$$f = a + bR$$

in which f is the critical frequency, R the relative sunspot number, and a and b constants dependent on the ionospheric layer, hour of day, time of year, and geographic location. (See Fig. 4.) Over a large range of values, such as have occurred during the present sunspot cycle, inclusion of a third term, cR^2, seems to be called for, c having a negative value. There is, of course, no physical reason for a linear (or any other definite) relationship between critical frequencies and the highly arbitrary sunspot number. The strongly linear tendencies of the relationship speak well of the intuition of the originator of the relative sunspot number system of measuring solar activity.

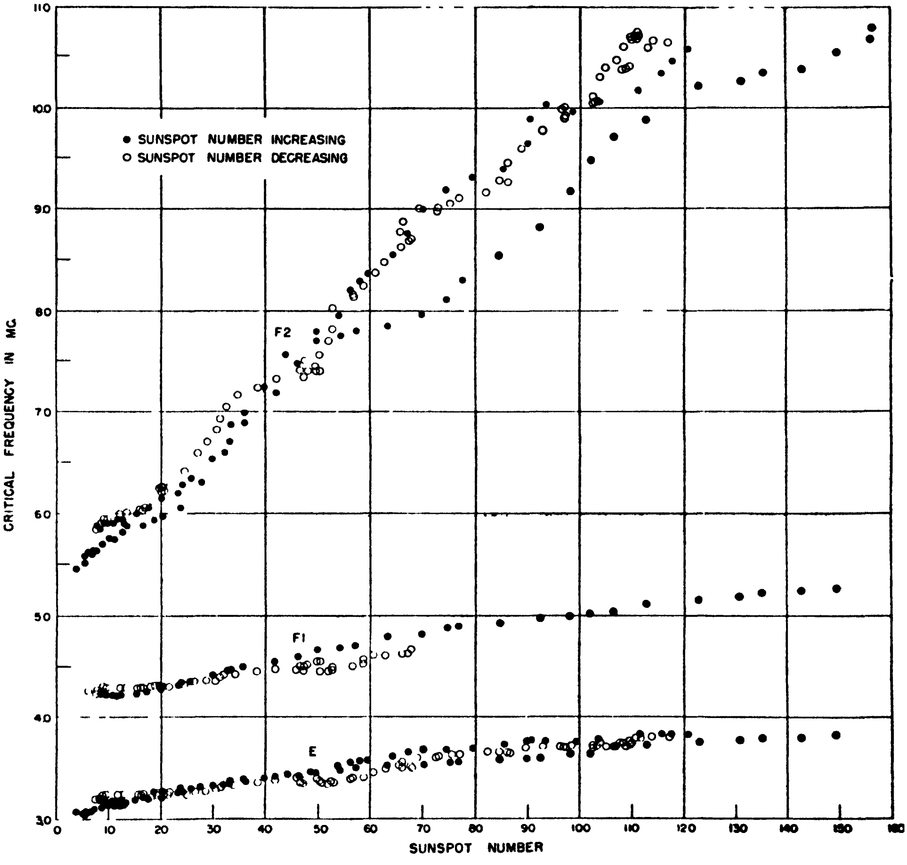


FIG. 4A.

FIG. 4.—Variation with sunspot number of critical frequencies at local noon of the E, F1, and F2 layers at (A) Washington, D. C., (B) Huancayo, Peru, and (C) Watheroo, Australia, 12-month running averages.

Since critical frequencies are proportional to the square root of ion densities the percentage change in ion density with sunspot number is even more pronounced. An estimate of the magnitude of the change in the solar ionizing radiation responsible for the several layers involves an assumption as to the processes of ionic equilibrium involved. If the process of electron removal is by attachment to neutral molecules, then the rate of removal will be proportional to the number of electrons present and a twofold increase in critical frequency implies a fourfold increase in

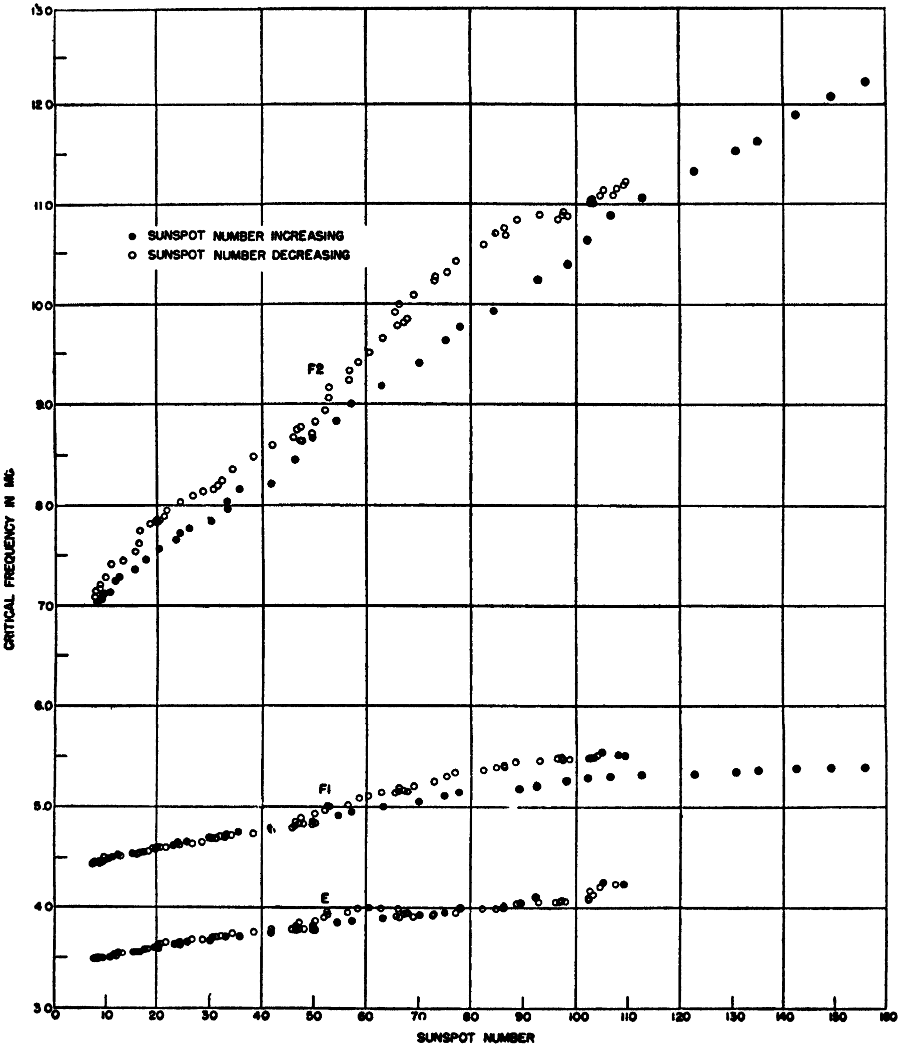


FIG. 4B.

the rate of ion production, for equilibrium conditions. If, on the other hand, electrons are removed by recombination with positive ions, and if the positive ions are numerically equal to the electrons, then for equilibrium conditions, a twofold increase in critical frequency implies an eightfold increase in ionizing radiation. It is not possible at the present time to decide which, if either, of these alternatives is correct. It is clear, however, that variation in the ionizing radiation from sunspot minimum to sunspot maximum is very great in spite of the high constancy of radiation in the visible spectrum. But the variation in the ionization radiation is not as great as the variation in sunspots, for even at the minimum of the sunspot cycle, when there may be no observable spots upon the sun, the ionizing radiation is sufficiently strong to support considerable ion densities.

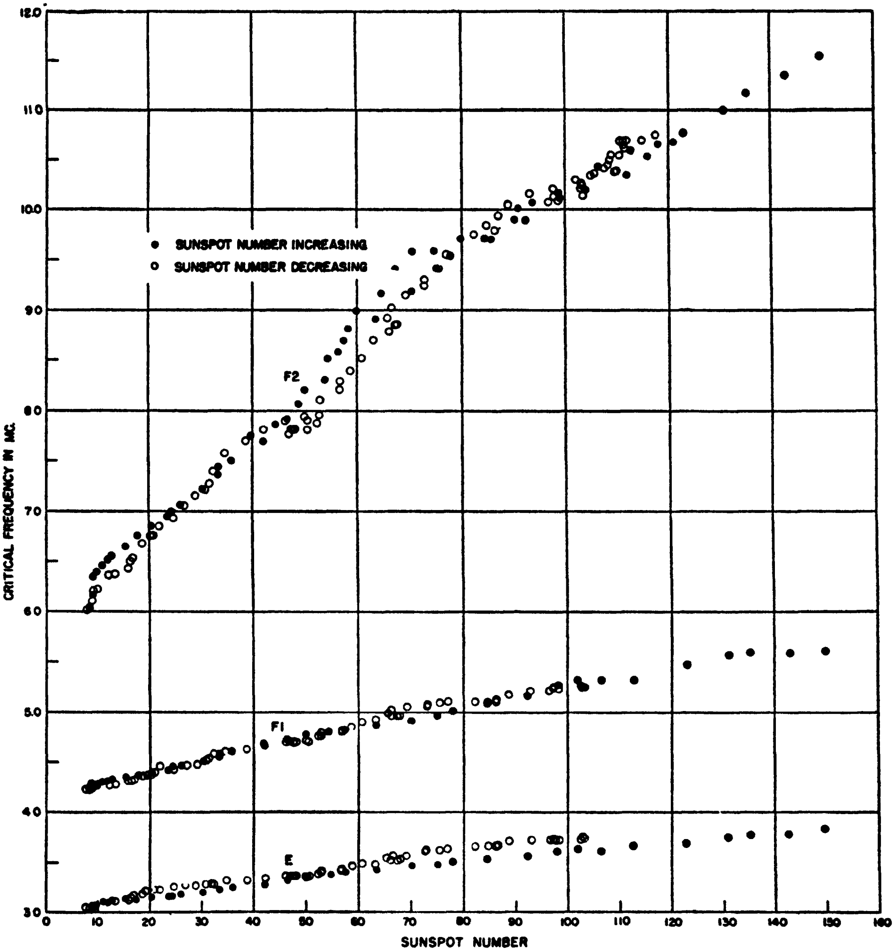


FIG. 4C.

The close relationship between critical frequencies and sunspot numbers suggests that critical frequencies may constitute a more satisfactory index of the 11-year change in the sun's behavior than the artificially contrived sunspot number based on counting the individual visible spots and groups of spots. An index of solar activity based on the critical frequencies eliminates the personal bias of the observer, a defect in the significance of sunspot numbers as a measure of solar activity, particularly when long-term trends are involved. On the other hand, critical frequencies exhibit seasonal and diurnal variations, but allowance can be made for these effects because of well-established trends.

By using the F2 critical frequency data for certain hours at Washington, Watheroo, and Huancayo, for which the seasonal trends have been well established it has been possible to establish an index of solar activity which closely parallels the variations of the relative sunspot numbers. Monthly values of this index exhibit smaller variations from their 12-month running averages than do corresponding values of the relative sunspot numbers, indicating that, in addition to being a more objective measure, it is also subject to smaller statistical fluctuations.

VII. Prediction of Ionospheric Disturbances

Associated with increases in solar activity are increases in the frequency of incidence of disturbances of the ionosphere. Since the effect of these disturbances is almost invariably a degradation of the ability of the ionosphere to sustain radio communication, the forecasting of disturbance became a matter of primary importance during the war when the success of military operations often depended on maintenance of reliable radio communication. During the war when conventional aids to navigation could not be employed and dependence had to be placed on long range radio aids, the probability of occurrence of ionospheric disturbances was considered in planning air raids.

The phenomena classified as ionospheric storms were recognized before the war as the ionospheric aspects of electromagnetic disturbances of the earth, other aspects of which had long been recognized in geomagnetic and earth-current storms and auroral displays. When the requirement for prediction of ionospheric storms became acute the abundant fund of information obtained from the lengthy series of data from geomagnetism, extending back for a hundred years, was available. Fortunately, the first years of the war fell during a part of the sunspot cycle for which the established 27-day recurrence tendency in geomagnetic disturbance was well marked. This recurrence tendency, due presumably to the 27-day rotation of the sun, in the course of which active areas are turned toward or away from the earth, is manifested in greater probability of the occurrence of a geomagnetic storm 27 days after one has occurred. On the basis of this recurrence tendency alone relatively reliable predictions could be made nearly a month in advance.

An intense program of observation of the visible solar phenomena was maintained in order to improve the forecasts by inclusion of such additional information as the solar observatories would supply. Regular observations of the sun's corona were instituted in order that active regions on the sun could be detected while still on the limb of the rotating sun. This involved employment of the recently developed coronagraph technique by means of which an "artificial eclipse" is produced by occulting the image of the bright disc of the sun.

The series of special solar observations are as yet too short to derive any reliable statistical conclusions relating visible solar phenomena with ionosphere storms. No single factor among the several solar characteristics observed has appeared as a reliable indicator of impending ionospheric disturbance. The presence of regions of unusual activity on the part of the sun which is turned toward the earth has served as a fairly good reason for the anticipation of disturbance.

If it is not required that warnings of ionospheric disturbance be given days in advance a rather high degree of success in their prediction can be achieved. During the initial stages of a geomagnetic storm the ionosphere is ordinarily not greatly affected. Therefore if the recordings of a magnetograph are closely monitored, advance warning of ionospheric disturbance may be given. The auroral regions of the earth are most sensitive to the effects of ionospheric disturbance, therefore considerable advance information on disturbance can be obtained by radio bearings on stations, the paths to which intercept the auroral zone, before the effects have become great enough to hamper ordinary communication. By application of these principles in conjunction with anticipation of disturbance from recurrence tendencies and solar phenomena a high degree of success can be achieved in warning of degraded propagation conditions for the following 24 hours. Such information is regularly broadcast at half-hour intervals by WWV, the standard frequency station of the Central Radio Propagation Laboratory, National Bureau of Standards.

A method for predicting sudden ionospheric disturbances on a probability basis was developed by statistical study. These disturbances are distinct from ionospheric storms. They consist of sudden increases in ion density in the absorbing region which persist for from several minutes to sometimes several hours. During these disturbances high frequency radio communication is seriously hampered or completely obliterated. Since the recognition of this special type of disturbance did not occur until 1935, and since the geomagnetic effect occasionally associated with the disturbance is not easily recognizable *per se* it has been possible to describe the statistical characteristics of the type only recently. No attempt is made to predict the exact time of occurrence of a sudden ionospheric disturbance; only the probability of the occurrence of one on a given day, based upon established recurrence tendencies and the presence of areas on the visible side of the sun that are recognized as typical sources of the bright chromospheric eruptions that cause the disturbances, is predicted. In spite of their brief duration these disturbances cause considerable interference with radio traffic, particularly when operation is automatic. Ability to anticipate their occurrence is thus of considerable value, especially around times of maximum solar activity when they are most numerous.

VIII. SPORADIC E REFLECTIONS

The elusive problems of sporadic E reflections, that is, occasional reflections from the E region at frequencies above the normal critical frequency of that region, have not been resolved by the accumulated data from the world-wide network of ionosphere stations. Because of differences in types of vertical incidence equipment in use and differences in their sensitivities, observations of sporadic E reflections at different stations are not readily comparable. Even at a single station, the diurnal, seasonal, and long-term variations in absorption together with changes in the performance of equipment confuse interpretation of the observations.

All sporadic E reflections do not seem attributable to the same cause. Several main types are widely recognized. In equatorial regions reflections from the E layer at frequencies above the E critical frequency are regular in their occurrence. The highest frequencies are reflected around noon. These reflections are ordinarily weak and the upper layers are not occulted. The reflections seem to result from the presence of a sharp boundary in the the E layer. Near the auroral zone strong reflections are frequently returned from the E region of sufficient strength to occult the upper layers. Many of these occurrences are associated with typical geomagnetic disturbances known as bays. In contrast with the equatorial phenomena the frequency of occurrence in high latitudes is greatest at night. In temperate regions both types of sporadic E reflections occur and apparently some additional types besides. No correlation has been noted between occurrence of sporadic E reflections and magnetic activity in temperate regions.

Assessment of the importance of sporadic E reflections to radio communication is difficult. Because of the highly local nature of the phenomena it is unlikely that "two-hop" propagation solely by sporadic E reflections occurs frequently. However there are numerous cases of "single-hop" propagation which find their only plausible explanation in sporadic E reflections, and it is entirely likely that much of the propagation at frequencies in excess of the maximum usable frequencies of the normal layers occurs in this way.

Considerable investigation has been carried out on the relative frequency of occurrence of sporadic E reflections on various radio frequencies at individual stations. Results from these studies are not subject to the same uncertainties as studies of seasonal, long-term, and geographic distribution. For any given station for any season of the year the percentage of time, P, that sporadic E reflections are obtained above a frequency f is given approximately by

$$\log P = a - bf$$

where a and b are constants dependent on the station and season. This relationship affords a basis for estimating the percentage of time that sporadic E reflections may be expected to support propagation of radio waves at frequencies in excess of the normal maximum usable frequencies. Since a is small the relationship may be represented approximately by

$$f \log (1/P) = \text{constant.}$$

IX. ABSORPTION OF RADIO WAVES

Calculation of the field intensities of radio waves propagated over given paths rests to an even greater extent on actual communication experience than the selection of frequencies for use over the path. The practical communication engineer usually solves the problem of what power is required for operation of a circuit by having sufficient power available for use so that the strength of the received signals will be as great as is conceivably necessary. Such methods are obviously not economical from the individual operator's viewpoint; nor efficient when considering the tremendous radio traffic which the ionosphere is called upon to support. Excessive use of power on one frequency to sustain communications over a given path creates interference for users of the same frequency over other paths, and thus imposes a pre-emptive limitation on use of the ionosphere.

The reason for reliance on practical experience is due to the absence in the past of a satisfactory generalized basis for calculating absorption over long paths. The theory of absorption of electromagnetic radiation in ionized media is very complete but the application of this theory for the complicated conditions encountered in the ionosphere is difficult. Also, many of the parameters necessary for numerical calculation of the absorption over actual paths on the basis of the physical theory are not known. Attempts to evaluate these parameters by observing field strengths of various transmitters over various paths have been fraught with difficulties. One of the principal uncertainties is involved in varying modes of propagation of the waves between transmitter and receiver which cause differences in absorption and in the angle of arrival, the latter involving differences in antenna response. The mode of propagation involved cannot be inferred except by elaborate experimental procedures at the receiver or transmitter, or detailed knowledge of the state of the ionosphere along the path. Similar difficulties are encountered in applying available propagation information to the practical problems of radio communication.

To attack the problems of ionospheric absorption more directly measurements of absorption at vertical, or near vertical, incidence have been undertaken on various frequencies. Two methods of measurement have been employed, one using the pulse technique and the other the cw technique. The former has the advantage of discriminating between reflections, partial or complete, from individual layers. The latter method is much simpler experimentally although the measurements are less simple of interpretation. If vertical incidence measurements of ion densities and heights of the various layers are also performed at the same station where absorption measurements are made, segregation of effects due to different layers can usually be accomplished. Such absorption measurements have been carried out for several years at Great Baddow, England (using the pulse technique), and at Washington, D. C. and several other localities (using the cw technique). Noon values of absorption have been obtained by the British since 1935. Special experiments on absorption using the pulse technique have been carried out for very short periods of time at a number of other localities.

Results obtained from these various investigations are not entirely in agreement, but the lengthier series of observations indicate clearly a number of important features regarding absorption. According to the simple theory of layer formation (in which the recombination coefficient is assumed to be invariant with height) the theoretical absorption of radio waves at vertical incidence should be proportional to $\cos^{3} \chi$. Most of the data seem adequately fitted by $\cos \chi$ (see Fig. 5), although observations on isolated days seem to call for a higher value of the exponent. These observations support the formulas which have been used for calculations of field intensity over long paths which involve a linear function of $\cos \chi$. Although intraseasonally at one station the observations may be systematized by a single $\cos \chi$ law, this may not be accomplished interseasonally. Distinct differences in absorption for the same values of $\cos \chi$ from one season to another have been revealed, the absorption during winter being greater on certain frequencies for an equal value of $\cos \chi$.

Effects of solar activity on absorption are revealed by observations extending over a number of years. From a sunspot number of zero to a sunspot number of 100 the absorption index at vertical incidence increases about 70%. This increase parallels the increase in the normal diurnal variations in geomagnetism, as would be expected since the geomagnetic variations, like radio wave absorption, are believed to be dependent on ion densities in the same region of the ionosphere. (See Fig. 6.) Thus the relationship between solar activity and absorption may be confidently generalized as applying not only for the period over which it has been observed but for past solar cycles as well. The close relationship between radio wave absorption and solar activity, as evidenced by sunspots, is demonstrated by comparing the average absorption over sequences of days during which the sunspot number increased and then decreased. The pattern of change in sunspot number during such sequences also appears in the absorption data.

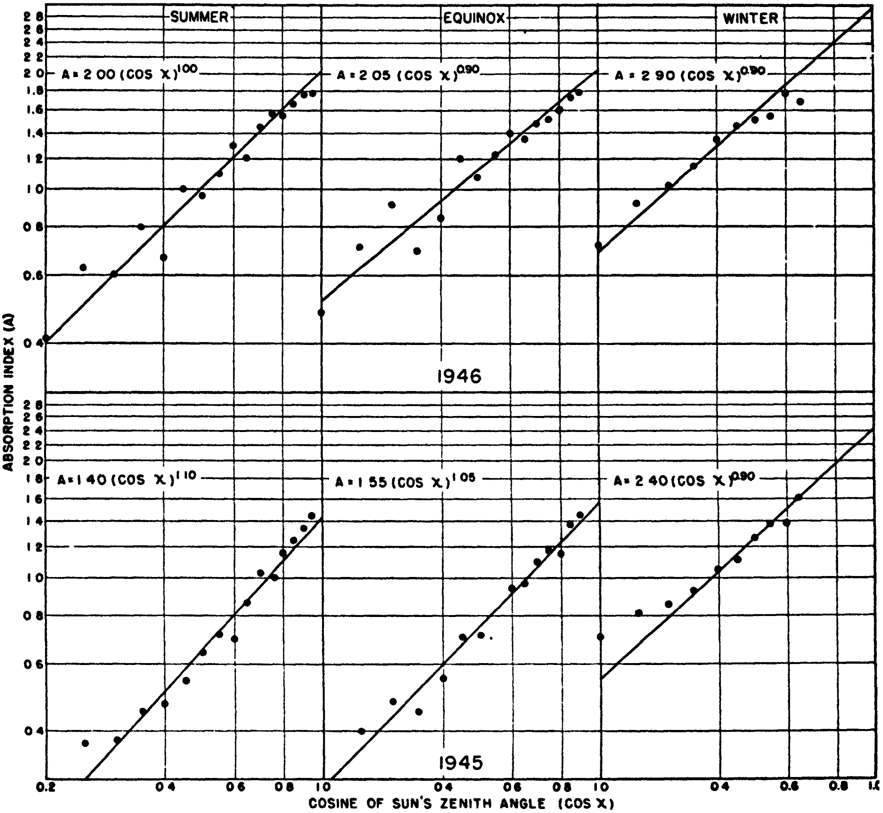


FIG. 5.—Variation of the absorption index at vertical incidence at Washington as a function of the sun's zenith angle.

While the actual change with solar activity in the absorption index at vertical incidence is small, its significance with respect to received field intensity at vertical incidence is large, and at oblique incidence, even larger. Let us consider a radio wave which is reflected from the F layer and which has an absorption index of 1.00 at vertical incidence at sunspot minimum. The equation relating field intensity to the absorption index is

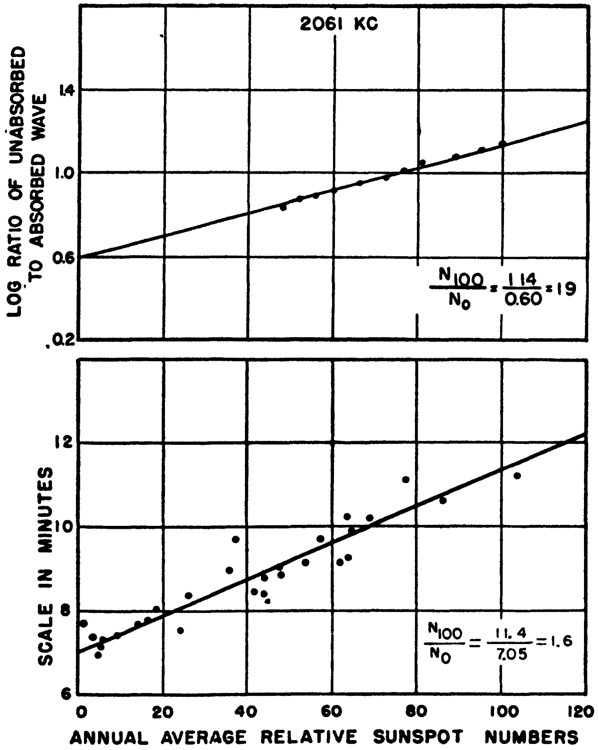


FIG. 6.—Variation with sunspot number of absorption index (A) and range in quiet-day diurnal variation in magnetic declination (B) in the vicinity of Washington, D. C.

$$E = E_0 \times 10^{-A}$$

where E_0 is the unabsorbed field intensity and A is the absorption index. Since the absorption index is the negative logarithm to the base 10 of a factor by which the field intensity in absence of absorption must be multiplied to allow for the decrease due to absorption, a 70% increase in the absorption index corresponds to a fivefold (14 decibel) decrease in field intensity. If the angle of incidences of the radio waves on the ionosphere is 60°, then the path length through the absorbing region will be twice as great and the absorption index will be doubled. Thus the absorption index over a path requiring an angle of 60° for reflection will increase from 2.0 to 3.4 for an increase in sunspot number from zero to 100, and the received field intensity will be reduced 25-fold (28 decibels). It might seem, in view of the foregoing, that oblique incidence observations would afford a more sensitive measure of the effect of solar activity on radio wave absorption. However, oblique incidence measurements are subject to the numerous uncertainties previously mentioned with respect to mode of propagation, which renders the interpretation of them difficult.

The belief that most ionospheric absorption occurs in the D region has had wide acceptance. However there are both theoretical and experimental reasons for supposing that if a radio wave penetrates the E layer, the absorption to which it is subjected in the E layer exceeds that to which it is subjected in the D region. The theoretical reasons for this is that the calculated product, $N\nu$ (N is electron density; ν electron collison frequency), upon which absorption depends, attains greater values in the E layer than in the D region. This involves an assumption regarding the level at which D region ionization becomes appreciable. Rocket sonde experiments indicate that this ionization does not become appreciable below 80 km. which is not contravened by radio observations. Using this height, the height and thickness of the E layer, as given in part by theory and in part by observations, and the values of electron density inferred from the radio observations, calculation of $N\nu$ and its integrated value with respect to height is a straight-forward process.

Direct experimental evidence that waves penetrating the E layer encounter their greatest absorption there is supplied by comparing the absorption at vertical incidence for different frequencies. Thus the absorption at Washington, D. C. measured during daylight hours of 2 Mc emissions is much less than the absorption of 4 Mc emissions when due allowance is made for the theoretical dependence of absorption on frequency. During the period when these measurements were obtained the 2 Mc emissions were reflected from the lower part of the E layer and the 4 Mc emissions penetrated the E layer and were reflected by the F1 or F2 layer. The difference in absorption can be quantitatively explained as a result of stronger absorption occurring in the E layer than in the D region.

So far little effort has been made toward quantitatively calculating absorption at oblique incidence from the vertical incidence absorption data. Practical calculations are based largely on empirical formulas which involve little implication regarding the exact levels at which absorption occurs. Vertical incidence studies have served to indicate how the formulas must be varied to allow for changes in absorption with solar activity, seasonal changes, and similar effects.

X. RADIO NOISE

If a radio signal is received at a certain place with a certain intensity, free of interferences from other radio signals, the usefulness of that signal for conveying intelligence is determined by the strength of the radio noise which is received with it. This noise may be either natural or man made.

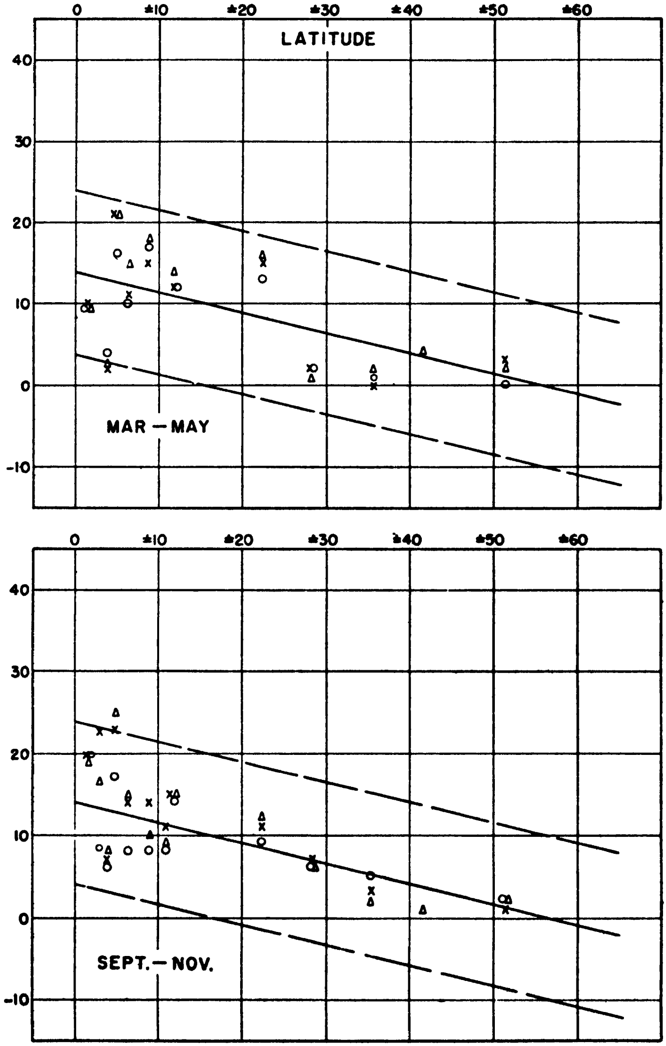
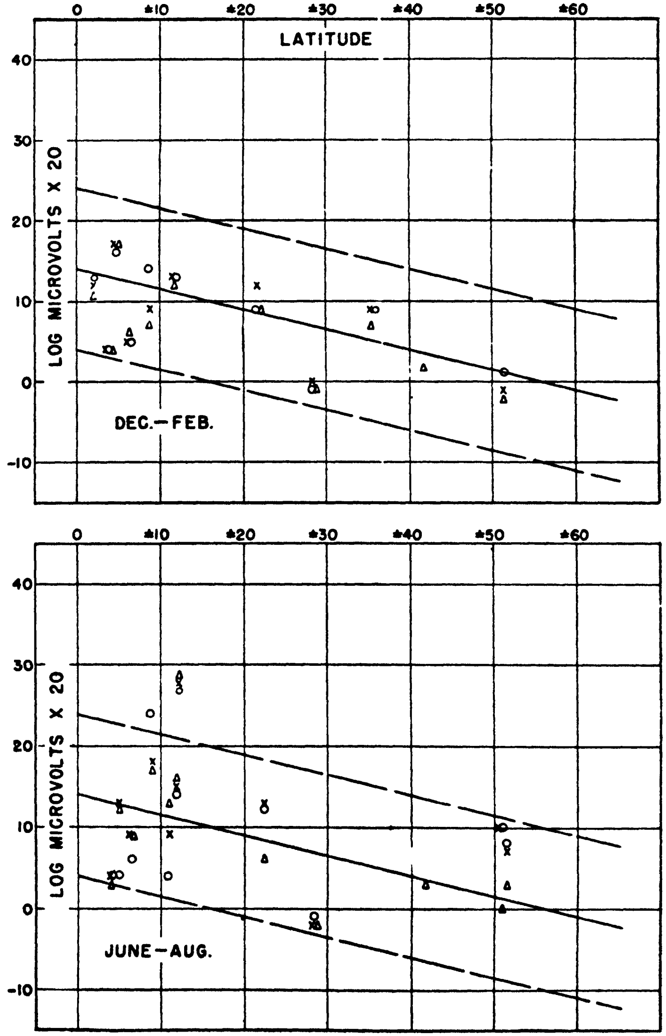


FIG. 7.—Latitude distribution of intensity of atmospheric radio noise at 5 Mc during

Good receiver design can always keep the internal noise of the equipment sufficiently low so as to be of no consequence. Man-made noise can usually be eliminated by choice of the receiving site and other devices, leaving the natural radio noise as the only uncontrollable element. Thus the study of natural radio noise is a vital phase of propagation research.

Thunderstorms are believed to be the major source of natural radio noise in the spectrum limits of waves propagated by the ionosphere, noise of cosmic or solar origin being of little consequence in comparison with



night hours exceeded 90 % of time (19h-21h, triangles; 23h-1h, crosses, 3h-5h, circles).

atmospheric noise. Thunderstorms in the immediate vicinity of the receiving site, while occasionally responsible for interruption of radio traffic, are, on the average, less important than the steady background of noise propagated by the ionosphere from distant sources. Thus it should be possible from knowledge of world-wide distribution of thunderstorms and ionosphere characteristics to predict the world-wide distributions of the intensity of radio noise.

At the beginning of the war there were only a few series of radio noise measurements, and these, in most cases, of only short duration. The requirement for estimates of the intensity of radio noise in various parts of the world was met by the construction of noise charts based upon accepted distributions of thunderstorms and knowledge of propagation conditions. In the absence of direct experience these charts furnished some basis for anticipating conditions when new regions were entered.

An observation program on radio noise was also launched and stations were established in many parts of the world using special equipment designed for the purpose. In some regions the observations agree closely with the conditions described by the charts, in others there are pronounced discrepancies. The reliance which can be placed on the observations is diminished by the limiting sensitivity of the equipment which is unable to determine the lower limit of noise values in many cases.

At most stations, atmospheric radio noise exhibits many similar characteristics. A pronounced diurnal variation is present nearly everywhere with lowest values around midday when noise being propagated from distant sources is strongly absorbed. On certain frequencies distinct increases in noise occur at the times and seasons when thunderstorms are prevalent for the region in which the noise station is located. Study of the night values of radio noise obtained by the network of noise stations, which are values upon which considerable reliance can be placed, show that the observed world-wide distribution of radio noise at this time is fitted better by assuming a linear decrease with increasing latitude than by the highly complex theoretical charts. (See Fig. 7.) A decrease in the night values of noise with increase in solar activity has also been noted, similar in magnitude to the increase in absorption with increase in solar activity. This may be attributable to the increased absorption of noise propagated from distant sources.

XI. Reflections from Meteor Trails

Interest in effects of meteors on the ionosphere existed prior to the war, but research in the field lagged until the cessation of hostilities in Europe. The investigations received fresh stimulation through chance observations of meteor trails by radar operators during the war. During the period when regular radar watches were no longer a prime necessity but the stations had not yet been de-activated, systematic watch was maintained for radar reflections from meteor trails and a large quantity of statistical data was compiled. Special investigations to inquire into the nature of the phenomena were undertaken.

Two techniques are employed for study of reflections from meteotrails, the observation of "whistles" and the observation of pulse reflecrtions. When a meteoric particle enters the atmosphere it is traveling at such a high velocity that it is capable of ionizing the molecules of the air. The ionization produced is so intense that for a short time it is capable of reflecting waves over a wide range of frequencies. If the emissions from a nearby radio station are being weakly received by ground wave, reflections from the meteoric ionization will produce interference with them. As the meteor shoots through the atmosphere the ray path to and from the advancing head of the column of ionization is subject to a change in length, giving a Doppler shift, which produces an audible frequency by interference. In case the direction and location of the meteor trail is such that the angles made by the incident and reflected rays to the trail are equal, the reflected signals are very strong. This is known as a "burst." If pulses are transmitted, instead of continuous waves, the time delay for receptions of the reflected pulse furnishes a direct measure of the distance to the ionized column. It has been suggested that pulse reflections are obtained only when conditions for a "burst" are fulfilled but many pulse reflections observed indicate this is not generally true. In some cases the time-delay has been observed to decrease rapidly during the initial moments of reflection, indicating a velocity of approach comparable with meteoric velocities. Presumably such reflections are returned directly from the advancing head of the column as for the case when whistles are observed.

The mechanism by which meteors produce ionization has not been established. Some ionization is probably produced by direct impact with the meteoric particle, some by collision with secondary high velocity particles, and some by photo-ionization. Geometric considerations suggest that the initial distribution of ion density in the column due to the second and third of these processes should be roughly proportional to the inverse distance from the axis of the column, outside of the small finite volume containing the axis, provided absorption of the ionizing agent is not too great. Decay of ion density probably results from two processes, diffusion, and recombination and attachment. Measurements of intensity and duration of reflection of radio pulses at various frequencies from the columns is thus a means for studying these processes in the ionosphere.

Reflections from meteor trails are obtained on frequencies in excess of 100 Mc which indicates that sometimes columns of appreciable cross-section have ion densities as great as 10^8 ions/cc. The number of trails capable of producing reflections is greater for lower frequencies, and the duration of reflections is also greater on the lower frequencies, if equal radiated power and receiver sensitivity are used. The precise law relating occurrence of reflections to the frequency of the emitted radiation has not been determined, but observations suggest that occurrence of reflections is inversely proportional to a power of the frequency higher than the first power. The duration of the echoes varies from a fraction of a second to a minute or more. In a few cases where radio observations are supplemented by visual observations, the duration of echo appears to be proportional to the visual magnitude of the meteor.

The possibility that sporadic E reflections are caused by meteors has been suggested. It is likely that many of the phenomena appearing on the multifrequency ionospheric records which are described as sporadic E may be caused by meteors However, reflections from meteor trails do not exhibit the characteristics revealed by many of the types of sporadic E reflections. Persistent strong propagation over long distances at frequencies well above the prevailing maximum usable frequency is not a condition which would be expected to result from the relatively short-lived and confined ionized trails of meteors. One propagation effect associated with meteors is occasional propagation of emissions from frequency modulation broadcast stations, over distances beyond the ground range, by meteor trails. Such "bursts," while often affording excellent reception momentarily, are of only very short duration.

XII. HIGH-SPEED MULTIFREQUENCY RECORDER

An outstanding development in equipment for ionospheric research occurred during the war. This is a high-speed multifrequency ionospheric recorder capable of sweeping through the frequency range from 1 Mc to 20 Mc in 7.5 seconds. Older automatic equipment requires from 1 to 20 minutes to cover a shorter range. With this new device, it is possible to obtain "pictures" of the virtual height-frequency curves for the ionosphere in sufficiently rapid succession that they may be shown on a screen by a moving-picture projector.

This new technique for studying the ionosphere has shown that many of the ionospheric changes occur with a rapidity which had not been anticipated from study of the more widely spaced older recordings. These effects are particularly pronounced during times of ionospheric disturbances. Also, the frequency sweep is now performed in a sufficiently short time that the record may be regarded as a fair representation of the instantaneous condition of the ionosphere at the time the record is made. When older methods of recording are used, appreciable changes sometimes occur in some layers of the ionosphere while the characteristics of other layers are being ascertained.

XIII. Trends of Research

While present trends in ionospheric research continue to place much emphasis on observational work, increased attention is being brought to the analysis and interpretation of the observations. Theoretical work during the war years suffered because of the lack of adequate observational material and because of the more urgent demands for practical applications.

The program of world-wide observation is being continued and extended to parts of the world that have previously been neglected. Particular attention is being given ionospheric research in very high latitudes where effects associated with the auroral zone render ionospheric problems much more complicated than they are in lower latitudes. Steps are being taken to bring procedures and techniques into coordination. Particular need for data on radio noise and absorption on a world-wide basis is recognized. In the interpretative aspects there exists a need for a scientific treatment of absorption phenomena and the development of a method for calculation of absorption without reliance on purely empirical formulas. This applies with great weight to absorption in the vicinity of the auroral zone, concerning which both scientific and empirical knowledge are extremely inadequate. Improved techniques for predicting ionospheric disturbances and their effects on communication are being sought.

Encompassing all, there is the need for a comprehensive theory of the ionosphere, its structure, and its variation, for observations during the war years have revealed an unsuspected complexity. This includes understanding of the physical properties and processes of the atmosphere, for the understanding of which the data obtained by rocket sondes and other technical developments of the war will prove of vital importance.

REFERENCES

In preparing a review of progress in a field of research during a war period great difficulty is attached to assignment of individual credit for specific advances. During the war years major attention was given to the exigencies of the military situation and much of the progress was due to hearty cooperation between organizations in individual countries. Therefore, reference to individual investigators has been avoided throughout the text. A list of selected references is given for those who may wish to pursue certain phases of the investigations in detail. Many additional investigations are described in reports of the various national organizations which carried out the research which has been described. Since these are not generally available to the public no reference has been made to them.

The first four references in the following list present general summaries of the field prior to the outbreak of war. The remaining references, listed in alphabetical order, include only papers written in English and published subsequent to 1938.

In addition to articles in scientific journals regular forecasts of radio propagation conditions six months in advance, prepared by the Central Radio Propagation Laboratory of the National Bureau of Standards, are published monthly in the United States by the Government Printing Office. Monthly data on the ionosphere collected by the world-wide network of ionospheric stations and other basic scientific data pertaining to the field are reproduced for limited circulation by the Laboratory to supply material for research workers.

1. Mitra, S. K. Report on the Present State of Our Knowledge of the Ionosphere. *Proc. Nat. Inst. Sci. India*, **1**, 131–215 (1935).
2. Mimno, H. R. The Physics of the Ionosphere. *Rev. Mod. Phys.*, **9**, 1–43 (1937).
3. Fleming, J. A., Editor. Terrestrial Magnetism and Electricity. McGraw-Hill, New York, 1939.
4. Chapman, S. and Bartels, J. Geomagnetism, Oxford University Press, Oxford, 1940.
5. Appleton, E. V. and Naismith, R. The Variation of Solar Ultraviolet Radiation During the Sunspot Cycle. *Phil. Mag.*, **27**, 144–148 (1939).
6. Appleton, E. V. and Benyon, W. J. G. The Application of Ionospheric Data to Communication Problems. Parts I and II. *Proc. Phys. Soc. Lond.*, **52**, 518–533 (1940); **59**, 58–76 (1947).
7. Appleton, E. V. and Naismith, R. Normal and Abnormal Region-E Ionization. *Proc. Phys. Soc. Lond.*, **52**, 402–415 (1940).
8. Appleton, E. V., Naismith, R., and Ingram, L. J. The Critical Frequency Method of Measuring Upper Atmospheric Ionization. *Proc. Phys. Soc. Lond.*, **51**, 81–92 (1939).
9. Bateman, R., McNish, A. G., and Pineo, V. C. Radar Observations during Meteor Shower. *Science*, **104**, 434–435 (1946).
10. Bates, D. R., Buckingham, R. A., Massey, H. S. W., and Unwin, J. J. Dissociation, Recombination and Attachment Processes in the Upper Atmosphere. *Proc. Roy. Soc.*, **170**, 322–340 (1939).
11. Berkner, L. V. Concerning the Nature of Radio Fade-out. *Phys. Rev.*, **55**, 536–544 (1939).
12. Berkner, L. V., Wells, H. W., and Seaton, S. L. Ionospheric Effects Associated with Magnetic Disturbance. *Terr. Magn. Atmos. Elect.*, **44**, 283–311 (1939).
13. Beynon, W. J. G. Oblique Radio Transmission in the Ionosphere and the Lorentz Polarization Term. *Proc. Phys. Soc. Lond.*, **59**, 97–107 (1947).
14. Booker, H. G. and Seaton, S. L. Relation between Actual and Virtual Ionospheric Heights. *Phys. Rev.*, **57**, 87–94 (1940).
15. Chamanlal, C. and Venkatamaran, K. Whistling Meteors—Doppler Effect Produced by Meteors Entering the Ionosphere. *Electrotechnics Bangalore*, **14**, 28–40 (1941).
16. Chapman, S. The Atmospheric Height Distribution of Band-absorbed Solar Radiation. *Proc. Phys. Soc. Lond.*, **51**, 93–109 (1939).
17. Cowling, T. G. The Electrical Conductivity of an Ionized Gas in a Magnetic Field, with Applications to the Solar Atmosphere and the Ionosphere. *Proc. Roy. Soc.*, **183A**, 453–479 (1945).

18. Darwin, C. The Refractive Index of an Ionized Medium. *Proc. Roy. Soc.*, **182**, 152–166 (1943).
19. Dellinger, J. H. Role of the Ionosphere in Radio Propagation. *Trans. Amer. Inst. Elect. Engrs.*, **58**, 803–822 (1939).
20. Eckersley, T. L. Analysis of the Effect of Scattering in Radio Transmission. *J. Inst. Elec. Engrs. London*, **86**, 548–567 (1940).
21. Eckersley, T. L. On the Existence of a Biannual Component in the F2 Layer. Ionization. *Terr. Magn. Atmos. Elect.*, **45**, 25–36 (1940).
22. Eckersley, T. L. and Farmer, F. T. Short Period Fluctuations in the Characteristics of Wireless Echoes from the Ionosphere. *Proc. Roy. Soc.*, **184**, 196–217 (1945).
23. Eckersley, T. L., Falloon, S., Farmer, F. T., and Agar, W. O. Wireless Propagation and the Reciprocity Law. *Nature, Lond.*, **145**, 222 (1940).
24. Ferraro, V. C. A. Diffusion of Ions in the Ionosphere. *Terr. Magn. Atmos. Elect.*, **50**, 215–222 (1945).
25. Harang, L. Scattering of Radio Waves from Great Virtual Distances. *Terr. Magn. Atmos. Elect.*, **50**, 287–296 (1945).
26. Harang, L. Polarization Studies of Echoes Reflected from the Abnormal E Layer Formed during Geomagnetic Storms. *Terr. Magn. Atmos. Elect.*, **46**, 279–282 (1941).
27. Hey, J. S. The Giacobinid Meteor Shower, 1946. *Nature, Lond.*, **159**, 119–121 (1947).
28. Hey, J. S. and Stewart, G. S. Derivation of Meteor Radiants by Radio Reflection Methods. *Nature, Lond.*, **158**, 481–482 (1946).
29. Hulburt, E. O. The E Region of the Ionosphere. *Phys. Rev.*, **55**, 639 (1939).
30. Huxley, L. G. H. The Propagation of Electromagnetic Waves in an Atmosphere Containing Free Electrons. *Phil. Mag.*, **27**, 313 (1940).
31. Manning, L. A., Helliwell, R. A., Villard, O. G., and Evans, W. E. On the Detection of Meteors by Radio. *Phys. Rev.*, **70**, 767–768 (1946).
32. Martyn, D. F. Anomalous Behavior of the F2 Region of the Ionosphere. *Nature, Lond.*, **155**, 363–364 (1945).
33. Mohler, F. L. Recombination and Electron Attachment in the F Layers of the Ionosphere. *J. Resarch Natl. Bur. Standards*, **25**, 507–518 (1940).
34. Pekeris, C. L. The Vertical Distribution of Ionization in the Upper Atmosphere. *Terr. Magn. Atmos. Elect.*, **42**, 205–211 (1940).
35. Phillips, M. L. Variations in Sporadic-E Ionization Observed at Washington, D.C. *Trans. Amer. Geophys. Union*, **28**, 71–78 (1947).
36. Phillips, M. L. The Ionosphere as a Measure of Solar Activity. *Terr. Magn. Atmos. Elect.*, **52**, 321–332 (1947).
37. Pierce, J. A. Ionization by Meteoric Bombardment. *Phys. Rev.*, **71**, 88–92 (1947).
38. Pierce, J. A., and Mimno, H. R. The Reception of Radio Echoes from Distant Ionospheric Irregularities. *Phys. Rev.*, **57**, 95–105 (1940).
39. Rydbeck, O. E. H. The Propagation of Electromagnetic Waves in an Ionized Medium, and the Calculation of the True Heights of the Ionized Layers of the Ionosphere. *Phil. Mag.*, **30**, 282–293 (1940).
40. Rydbeck, O. E. H. The Reflection of Electromagnetic Waves from a Parabolic Friction-free Ionized Layer. *J. Appl. Phys.*, **13**, 577–581 (1942).
41. Shultz, E. L. Comparison of Predictions of Atmospheric Radio Noise with Observed Noise Levels. *Trans. Amer. Geophys. Union*, **26**, 854–860 (1947).

42. Smith, N. The Relation of Radio Skywave Transmission to Ionospheric Measurements. *Proc. Inst. Radio Engrs.*, **27,** 332–347 (1939).

43. Smith, N. Oblique Incidence Radio Transmission and the Lorentz Polarization Term. *J. Research Natl. Bur. Standards*, **26,** 105–116 (1941).

44. Stewart, J. Q., Ference, M., Slattery, J. J. and Zahl, H. A. Radar Observations of the Draconids. *Sky and Telescope*, **6,** 3–5 (Mar. 1947).

45. Wells, H. W. Sporadic E-region Ionization at the Watheroo Magnetic Observatory. *Trans. Amer. Geophys. Union*, **26,** 381–387 (1945).

46. Wells, H. W. Effects of Solar Activity on the Ionosphere and Radio Communications. *Proc. Inst. Radio Engrs.*, **31,** 147–157 (1943).

47. White, F. W. G. The Estimation of Wireless Transmission Data from Ionospheric Observations. *N. Z. J. Sci. Tech.*, **21,** 114–127 (1939).

48. White, F. W. G. The dispersion of Wireless Echoes from the Ionosphere. *Proc. Phys. Soc. Lond.*, **51,** 859–864 (1939).

49. White, F. W. G. and Straker, T. W. The Diurnal Variation of the Absorption of Wireless Waves. *Proc. Phys. Soc. Lond.*, **51,** 865–875 (1939).

Cosmic Radio Noise

JACK W. HERBSTREIT

Central Radio Propagation Laboratory
National Bureau of Standards
Washington, D. C.

CONTENTS

		Page
I.	Introduction	347
II	Jansky's Measurements	348
III	Reber's Early Measurements	349
IV	Later Measurements	354
V	The Point Source in Cygnus	356
VI	National Bureau of Standards Measurements	356
VII.	Method of Measurement	358
VIII	Results of Measurements	364
IX	Analysis in Terms of External Noise Factors	366
X.	Field Intensities Required for Communication Services	369
XI.	Effective Temperature Concept	369
XII.	Distribution of the Intensity of the Noise Sources with Direction and Frequency	370
XIII	Intensity from Small Noise Sources .	375
XIV	Observed Intensity of Radio Frequency Radiation from the Sun .	376
XV	Polarization of Extraterrestrial Radiation	378
XVI	Origin of Cosmic Radio Noise	379
	References	380

I. INTRODUCTION

It has long been recognized in all systems of communications that a requirement for the reception of intelligence from a desired signal is that it must be strong enough to be detected in the presence of interference either from undesired communication signals or from noise originating in a variety of sources both internal and external to the receiving apparatus. One type of interference to radio communications originating within the receiver is the random noise due to the fluctuation of electrons either in the resistance components of impedance elements or in vacuum tubes.

A well recognized type of external interference originating in the earth's atmosphere is the radio noise associated with the discharge of lightning in thunderstorms and propagated to the receiving location over the surface of the earth and via the ionosphere.

It has only been quite recently that ever present energy in the form of external random noise was discovered to be arriving at the earth in all directions from outer space. It is this extraterrestrial radio frequency energy that has been termed "cosmic" or by some recent experimenters "galactic" radio noise. The existence and extraterrestrial origin of this type of noise was first recognized by Karl Jansky of the Bell Telephone Laboratories in 1932 while he was measuring the directional properties of atmospheric noise at 14.6 meters (20.5 Mc) at the Bell System Holmdel, N. J. Laboratories.[1] Since noise is the factor which limits the detectability of radio communication signals, it is important that the characteristics of cosmic noise be determined in regard to its directional properties, absolute magnitude, and frequency dependence. In addition, these determinations will undoubtedly provide valuable information regarding the nature of our universe.

II. JANSKY'S MEASUREMENTS

Jansky noticed that the average level of the background, hiss type of noise received at 20.5 Mc changed as he rotated a highly directional receiving antenna. At first he thought that this change in received noise level was due to an unidentified carrier modulated by the noise in his receiver. However, after recordings were made over a short period of time, he noticed that the noise seemed to come from the direction of the sun. When the records made over a period of a year had been analyzed, it turned out that the direction of arrival of the maximum noise changed from month to month and at the end of a year had returned to approximately the same position as at the start.[2] This immediately suggested to him that the main source of this noise was not primarily the sun but that it originated elsewhere in outer space. Detailed analysis of the direction of arrival of the noise indicated the position of the apparent source of this noise as being in the region of the constellation Sagittarius, the approximate celestial coordinates of which are thought to be the center of our galaxy. The fact that the noise Jansky noticed appeared to come from the sun was merely a coincidence in that the sun happened to have approximately the same celestial coordinates as the center of the galaxy at the time of his original observations. The antenna used by Jansky was primarily directional in the horizontal plane and could be rotated through 360° of azimuth, having a beam width to half power response approximately 37° in height and 30° in azimuth.

In 1937, further experiments were conducted by Jansky at 18 Mc using a fixed rhombic antenna directed northeast and southwest, and at

9.3, 18, and 21.4 Mc using half-wave dipole antennas having very broad directional characteristics over the sky.[3] These measurements gave further evidence pointing to the fact that radio frequency energy of extraterrestrial origin is arriving at the earth. With the half-wave dipole antennas, Jansky attempted to determine the way in which the cosmic noise varied with frequency; however, at the relatively low frequencies used, ionospheric absorption was an appreciable factor, being most noticeable at the lowest frequency, so that a conclusive frequency law was not obtained.

III. REBER'S EARLY MEASUREMENTS

In 1940, Grote Reber, working in his laboratory at his home in Wheaton, Ill., constructed a unique tuned-radio-frequency amplifier receiver, using acorn tubes and transmission line tuned circuits, for receiving cosmic radio noise at 160 Mc.[4] Fig. 1 shows Reber's directional antenna or energy collector used for his measurements. It consists of a large parabolic sheet metal mirror with a half-wave dipole at the focal point. The dipole and receiver are mounted at the upper right edge of the photograph. The entire mirror was constructed on tracks which permitted rotation of the antenna in a north-south plane to any desired celestial declination angle almost down to his horizon at approximately −32° celestial latitude. The normal daily rotation of the earth provided the additional rotation necessary for observing the intensity of cosmic radio noise within his field of view. The directional characteristics of the antenna were such that it responded primarily to energy arriving in a narrow cone approximately 14° in diameter. This relatively narrow beam width permitted plotting contours of constant cosmic radio noise level over that portion of the celestial sphere visible at Wheaton, Ill., thus locating the sources of cosmic noise in some detail. His results confirmed Jansky's previous lower frequency observations that the main source of the noise was in the region of the constellation Sagittarius at celestial coordinates of 17 hours 50 minutes right ascension and declination of −25°. Reber also located a secondary maximum in the region of the constellation Cygnus,[5] having previously deduced from observations of Friis and Feldman[6] that this maximum existed. The contour charts prepared from his measurements at 160 Mc are shown in Fig. 2.

Fig. 3 is a world star map and shows the location of these constellations in the celestial sphere. The dashed line is the ecliptic or the apparent path of the sun in the celestial sphere. At the vernal equinox, approximately March 21, the sun is at zero hours right ascension and 0° declination. At the summer solstice, the sun is at 6 hours right ascension and declination approximately +23½°. At the autumnal equinox, September 21, the sun is at 12 hours right ascension and 0° declination, and at the winter solstice, December 21, at 18 hours right ascension and −23½° declination.

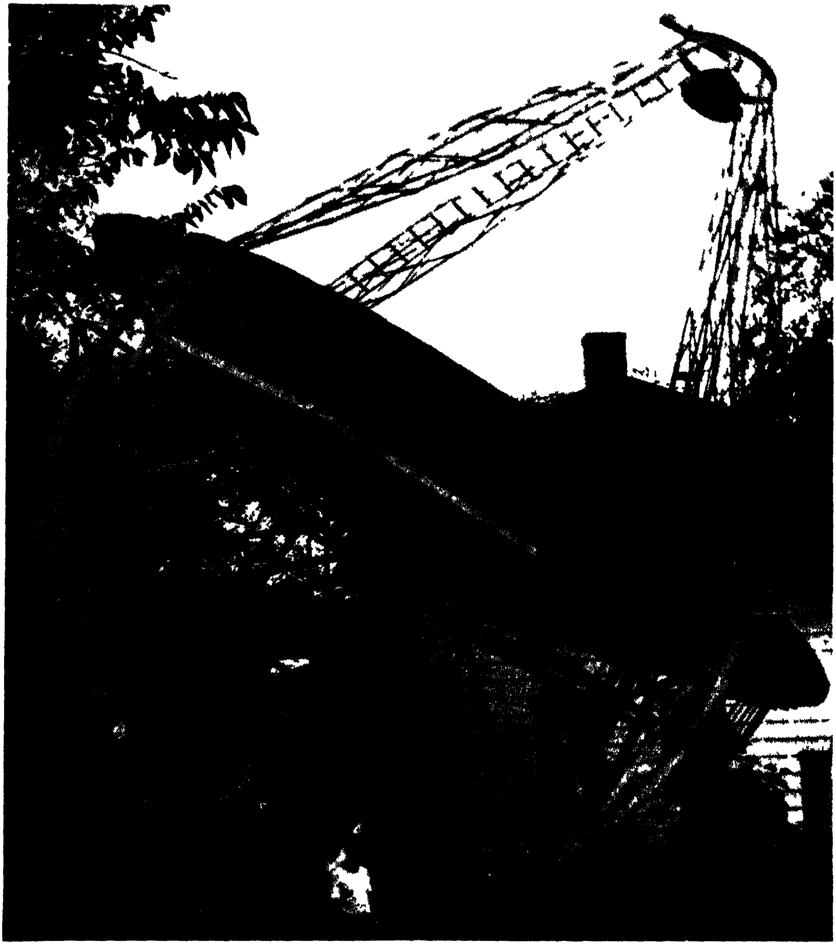
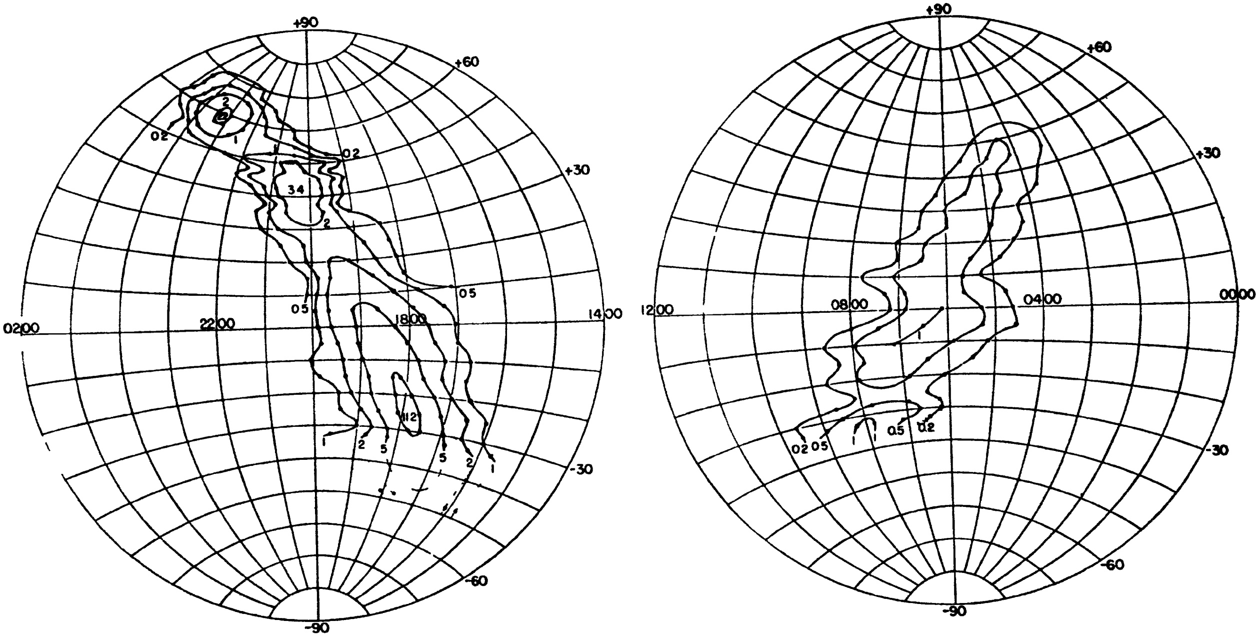


FIG. 1.—Large directional antenna used by Reber to receive cosmic radio energy. (Courtesy of Grote Reber.)

Jansky's very first observations were made at approximately the winter solstice with the sun at the same celestial coordinates as Sagittarius so that the assumption that the extraterrestrial noise was of solar origin was an easy one to make.

In 1943, Reber with his higher gain and more directive antenna did detect and record noise of solar origin in addition to the cosmic noise on



Unit on contours = 4.8×10^{-22} watts per square meter per steradian for a one cycle bandwidth

FIG. 2.—Contours of constant received cosmic radio noise plotted by Reber at 160 Mc. (Courtesy of Grote Reber.)

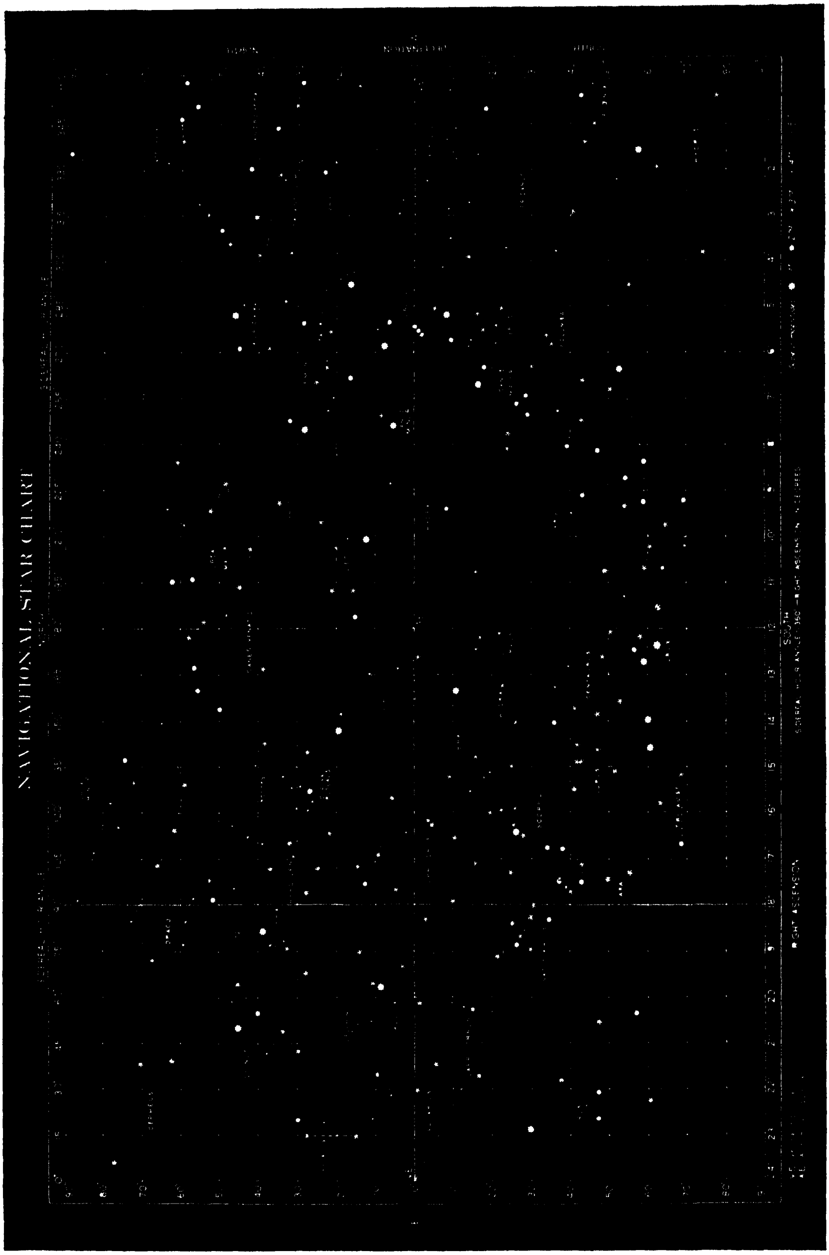


FIG. 3.—Navigational star chart.

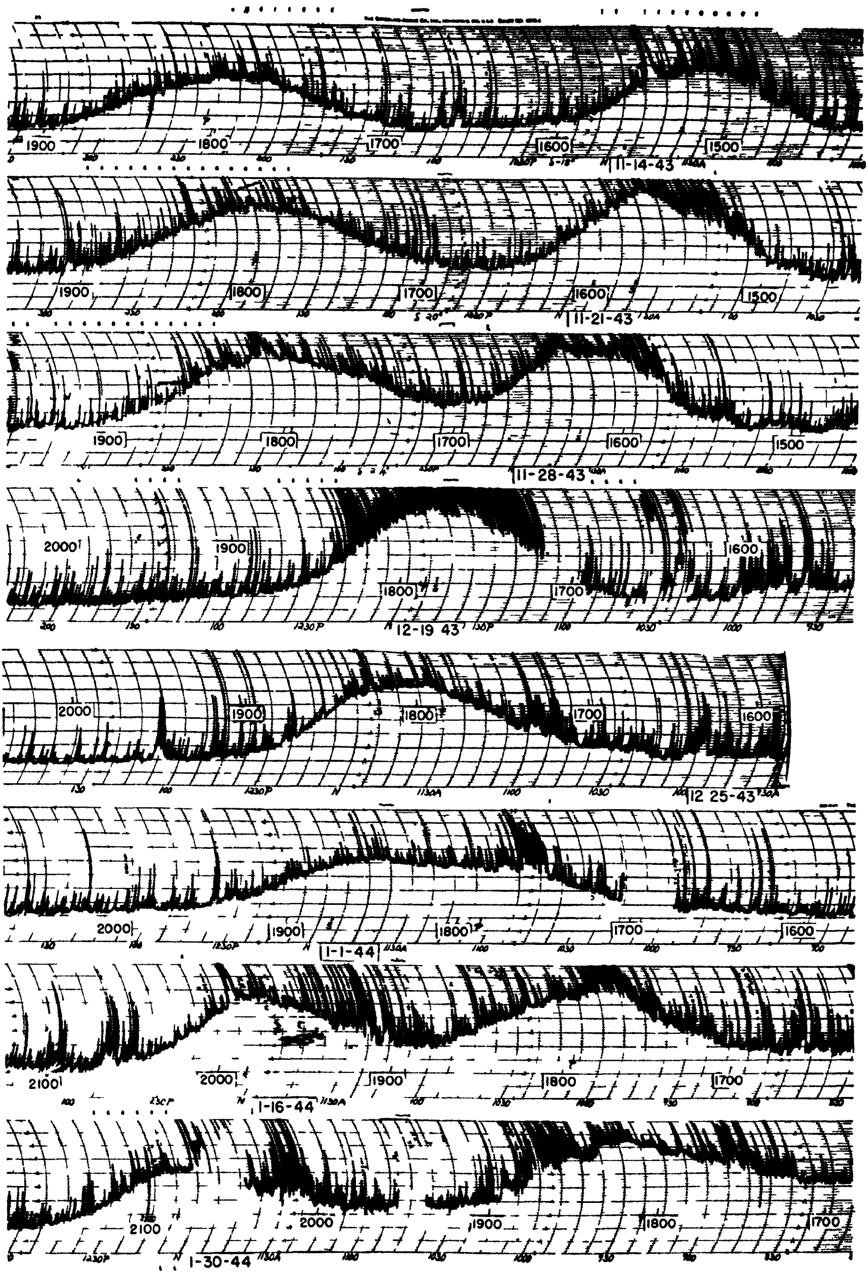


FIG. 4.—Cosmic and sun noise recorded by Reber 1943–1944, 160 Mc. (Courtesy of Grote Reber.)

160 Mc.[7] A portion of his records taken around the winter solstice are shown in Fig. 4 and clearly illustrate the reason for Jansky's first conclusion. These charts were all made with the antenna pointed at a fixed declination angle, the rotation of the earth being responsible for the change in the area of the sky at which the antenna was pointed. Both right ascension and local times are shown on each chart. In the top chart, the first maximum occurred when the antenna was pointed at approximately 15 hours right ascension, the position of the sun along the ecliptic at that time. The second maximum was due to cosmic radio noise and occurred when the antenna was pointed toward the region of Sagittarius at 18 hours right ascension. As the year progressed, in other words as the sun moved along the ecliptic, the two maxima are seen to appear closer together and in the region of the winter solstice, they coincide as shown in the fourth chart. The succeeding charts show the sun moving out from the center of the large cosmic radio noise source as it moves further along the ecliptic.

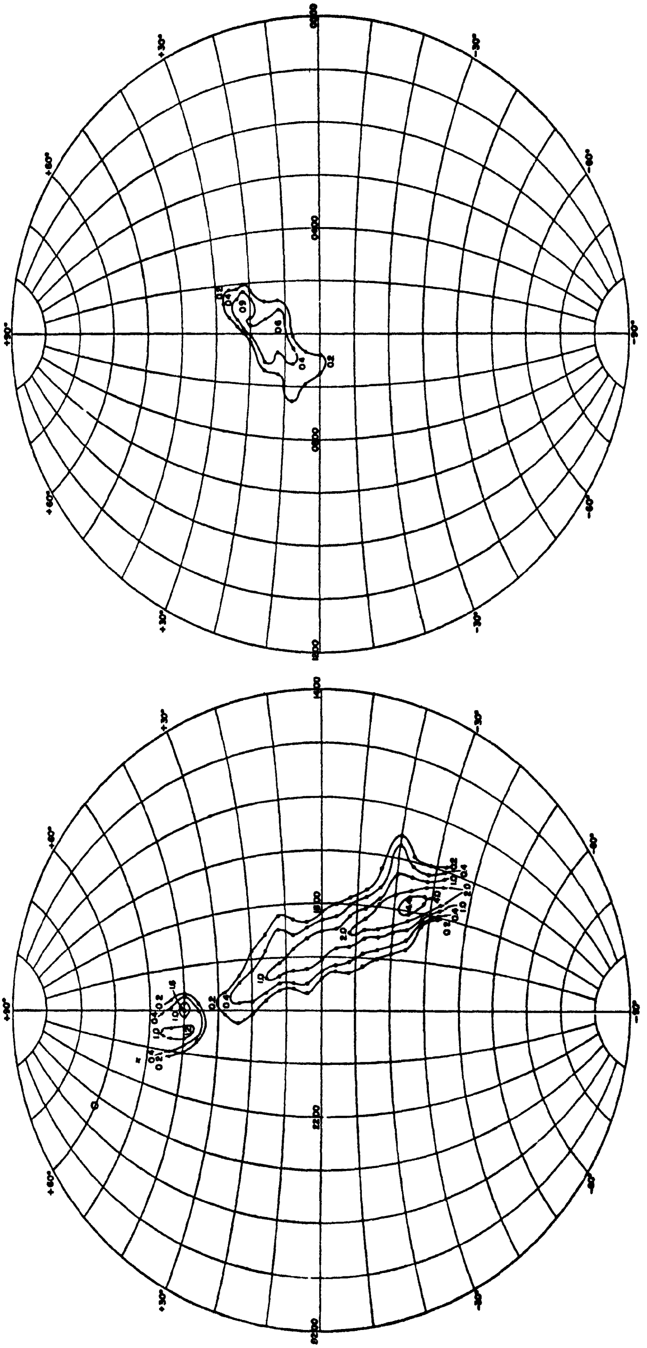
IV. LATER MEASUREMENTS

In 1945, Hey, Parsons, and Phillips[8] measured the intensity of cosmic radio noise at 64 Mc in England and prepared constant intensity contours of the noise which were similar to those plotted by Reber in 1944 for 160 Mc. These measurements were made with a small Yagi antenna array with a beam width to half power response approximately 12° wide in elevation and 30° in azimuth. The maximum of the beam was at a fixed elevation of 12° and could be rotated to any bearing.

K. F. Sander of the Radar Research and Development Establishment in England measured the intensity of cosmic noise at 60 Mc in the spring of 1945 using an antenna array consisting of four colinear horizontal half wave dipoles in front of a reflecting screen.[9] The antenna beam width was approximately 20° in azimuth and 30° in elevation, with the maximum response approximately 30° above the horizon.

Also in 1945, L. A. Moxon of the Admiralty Signals Establishment in England made measurements of cosmic noise at 90 Mc using modified radar equipment having an antenna with a beam width approximately 45° wide at half power response.[10] Contours of noise intensity over the sky were not plotted from either of these measurements; however, the diurnal patterns obtained were as would be expected from a major source in the Milky Way as found by the other investigators. The absolute values measured also are an aid in determining the frequency law of the noise.

In 1946, using more directive antennas, Hey, Parsons, and Phillips reported[11] observing the noise from a small area in the region of Cygnus



Unit on contours = 8.1 X 10^{-22} watts per square meter per steradian for a one cycle bandwidth

FIG. 5.—Contours of constant received cosmic radio noise plotted by Reber at 480 Mc. (Courtesy of Grote Reber.)

to be of varying intensity as it rose and set through their antenna beam. The antenna used for these observations was approximately 12° in diameter with maximum response fixed at 12° above the horizon and variable in azimuth.

In 1946 and 1947 Reber made a study of his visible sky at 480 Mc using the same parabolic sheet metal mirror for collecting the cosmic radio energy as was used on 160 Mc. At this much higher frequency the antenna, having the same area, had a narrower beam width, approximately 5° in diameter, which provided a means for determining the source of the noise in still greater detail. The results of Rebers measurements on 480 Mc are shown in Fig. 5. These data show two maxima in the region of Cygnus.

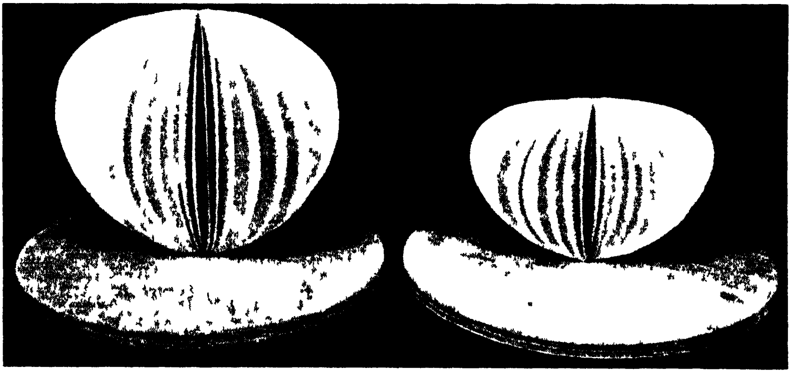
V. The Point Source in Cygnus

During a recent visit to the United States, J. L. Pawsey of the Radio Physics Laboratory, Sydney, Australia has described new experimental work being done by J. G. Bolton and G. J. Stanley in Australia on the small source of extraterrestrial noise in Cygnus. Observations were made on three frequencies between 60 and 200 Mc using directional antennas located on cliff sites over-looking the sea. The resultant antenna directivity was an interference pattern consisting of a series of lobes with the maximum response occurring at angles above the horizon where the direct wave and the almost-perfectly-reflected wave from the surface of the sea arrived at the receiving antenna in phase. The source of noise was then observed as it rose over the horizon and passed through the lobes of the interference pattern. Using this method, the small source in Cygnus was determined to be less than about 8 minutes in diameter at 19 h 58′ 47″ ± 10″ right ascension and 41° 47′ ± 4′ declination. The received noise at the low frequency is found to be of varying intensity while at the higher frequency the variations are not observed. It has been suggested that ionospheric refraction and absorption, which would be expected to be most predominant at the lower frequencies, causes the observed variations in intensity just as atmospheric refraction causes twinkling of stars at light frequencies. Since Hey, Parsons, and Phillips observations of the source in Cygnus were made at a relatively low frequency this explanation may also apply to their results. The striking thing about these observations is that the apparent small source of the noise is from a very ordinary portion of the Milky Way in which no outstanding star source appears to exist.

VI. National Bureau of Standards Measurements

In the measurements outlined above, most of the research has been devoted to determining the spatial distribution of the sources of the noise using antennas having a variety of directivity patterns on frequencies ranging from approximately 9 to 480 Mc. Since different antenna patterns mean that different areas of the sky are being observed, these measurements do not provide comparable results as a function of frequency.

At the present time the National Bureau of Standards, Washington, D. C. is in the process of implementing a program of measurement of the intensity and frequency distribution of cosmic radio noise from approximately 25 to 160 Mc. In making these measurements, use is being made of antennas with comparable directivities and calibrating methods that are as nearly identical as possible for the recorders on each of the various frequencies. Up to the present time, measurements have been made on the frequencies 25 and 110 Mc using half-wave dipoles placed one-quarter wavelength above the ground. The directivity pattern of such an antenna over a perfectly conducting earth is a very broad elliptical cone approximately 100° wide to half power response in the plane normal to the dipole, and 60° wide in the plane of the dipole. When placed above an imperfectly conducting earth, the space pattern is modified by ground absorption of a portion of the incident energy. Fig. 6 shows solid models of directivity patterns for a 25 Mc horizontal half-wave dipole antenna receiving randomly polarized energy when located one-quarter wavelength above (a) a perfectly conducting earth and (b) ground with conductivity of 0.002 mhos/meter and dielectric constant of 4. Measurements have been obtained using antennas of this type on 25 and 110 Mc



(a) (b)

FIG 6—Half-wave dipole antenna directivity patterns at 25 Mc: (a) over a perfect ground, (b) over ground of conductivity σ = 0 002 mhos/meter and dielectric constant $\epsilon = 4$.

which provide the basis for a preliminary quantitative estimate of a frequency law for cosmic noise.

VII. Method of Measurement

Before proceeding further with a discussion of these results, the method of measurement will be described. During the war, considerable progress was made in understanding the nature of the limitations on very high frequency (VHF) and ultra high frequency (UHF) receiver sensitivity arising from the presence of various random noise sources within the receiver, due to the fluctuation of electrons in the resistance components of impedance elements and in vacuum tubes. In connection with this work, methods have been devised for measuring the absolute level of this noise in a radio receiver and thereby determining the absolute sensitivity of the receiver, for in the absence of external noise, it is the receiver noise alone that the signal has to compete with in order that it may be detected. Cosmic radio noise has been observed to have the same general properties as the fluctuation noise in the receiver. In other words, when we listen with headphones to a sensitive receiver with its directional antenna first pointed at open sky and then turned toward a source of cosmic radio noise, what is heard is an apparent increase in the receiver noise. This being the case, the methods of measurement of receiver noise are applicable to the measurement of cosmic radio noise.

The concept of "noise figure"[12] is now widely used as a measure of set noise and the absolute sensitivity of a radio receiver. To understand the meaning of the term "noise figure," the concept of *available power* must first be appreciated. The available power from a voltage generator is the power that it would deliver to a load with resistance equal to the internal resistance of the generator or, in other words, the maximum power that can be obtained from the generator is that which it will deliver to a matched load. If a load of different impedance is substituted, the actual power delivered would change, but since the generator has not been touched, the available power is unaltered. This simple concept allows us to separate the effects of load impedance from the properties of the rest of the circuit. The noise figure $\overline{NF}$ of a receiver in terms of available powers is defined as the ratio of the available signal power at the input to kTB, the available noise power from the passive resistance of the dummy antenna, divided by available signal to noise ratio in the output of the receiver.

$$\overline{NF} = \frac{\text{Available signal power at input}/kTB}{\text{Available signal power at output/Available noise power at output}} \quad (1)$$

where $k = 1.37 \times 10^{-23}$ = Boltzmann's constant
T = absolute temperature in degrees Kelvin (taken as 300°)

B = Effective noise band width of the receiver in cycles/second
$kTB = 4.11 \times 10^{-21}$ watts for a dummy antenna at room temperature ($T = 300°$) for a receiver noise band $B = 1$ cycle/second and represents the available noise power from the passive resistance of the dummy antenna.

There exists available noise power in any resistance because of the random motion of the electrons in the resistance, the absolute value of the available noise power being equal to kTB as derived by Nyquist in 1928.[18] Simply stated, the noise figure, $\overline{NF}$, is the ratio of the signal-to-noise power ratio at the input divided by the signal-to-noise power ratio at the output. We see that if we have the same available signal-to-noise ratio at the output as we have at the input, we would have a noise figure of 1 and a perfect receiver. In practice, the available signal-to-noise ratio at the output is lower than at the input because of noise added in the receiver and consequently the noise figure is greater than 1. At frequencies up to approximately 40 or 50 Mc, receivers have been built with noise figures approaching 1. However, as the frequency is increased, it becomes more and more difficult to achieve a low noise figure receiver. The variation of noise figure vs. frequency as measured for lighthouse tube radio frequency amplifiers in special low noise grounded grid circuits will be shown later.

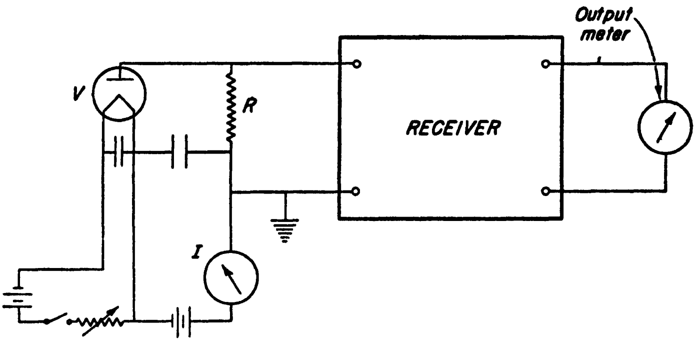


FIG. 7.—Noise diode method of $\overline{NF}$ measurement.

A method of measuring the noise figure of a receiver by means of a tungsten filament diode noise current generator used as a signal generator is shown in Fig. 7. When a diode is operated such that the plate current is limited by the temperature of the cathode, it has been shown that the noise current squared flowing in the diode is equal to twice the electronic charge, e, in coulombs, times the plate current, I, in amperes times the bandwidth, B, in cycles/second. Then the available power from the diode generator may be shown to be equal to $eIBR/2$, where R is the diode load resistance in ohms.

In making the measurements the diode is first connected to the receiver as shown with the diode load impedance equal to the antenna transmission line impedance. The diode plate current is next adjusted by varying the filament voltage and temperature so that the output power of the receiver as measured on the power output meter is just double the power output obtained with zero diode current. When it is considered that the output noise power contributed by the diode is actually the signal power, it is seen that the output has been adjusted for a signal-to-noise ratio equal to one. From the definition of noise figure, it is seen that, having made the output signal-to-noise ratio equal to one, the noise figure is equal to the ratio of the available signal power at the input to kTB. This ratio becomes simply $19.3IR$ at $T = 300$°K. when the values of e, k and T are substituted. The use of the diode noise source for measuring the receiver noise figure has the advantage that measurements of the receiver noise band width, B, are unnecessary.

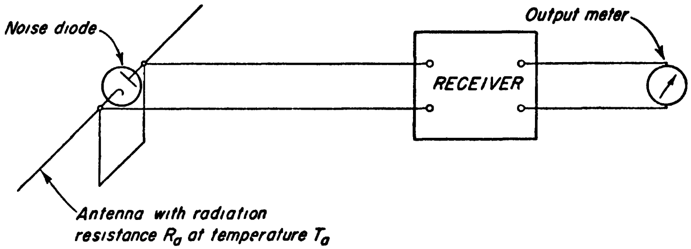

Noise diode
Output meter
RECEIVER
Antenna with radiation resistance R_a at temperature T_a

FIG. 8.—Measurement of external noise.

When the diode measurements are made of the noise figure of a receiver coupled to an actual receiving antenna of the same impedance as the diode load impedance used in the measurement of noise figure, then these measurements will determine an effective noise figure, $\overline{NF}'$, which may be either larger or smaller than the noise figure measured with the diode and real resistance load; this difference is due to the external noise picked up by the antenna. It should be noted that a radiation resistance is not a real resistor and thus introduces no noise into the receiver except to the extent that it absorbs noise radiation from its surroundings. It is convenient to express the available noise power picked up by the antenna, N_a, as equal to an external noise factor, $\overline{EN}$,[18] times kTB:

$$N_a \equiv \overline{EN}kTB \text{ watts } (T = 300^\circ) \tag{2}$$

Thus the dimensionless external noise factor, $\overline{EN}$, as defined by this relation, is a convenient measure of the external noise energy. $\overline{EN}$ may be determined by locating a calibrating diode right at the antenna terminals as illustrated in Fig. 8, and determining the effective noise figure $\overline{NF'}$ at the antenna terminals using a similar procedure to that used for determining $\overline{NF}$. Now, instead of determining the signal generator power to kTB power ratio at the antenna, the signal generator power to external noise power ratio is determined. $\overline{EN}$ is then obtained from the following relation:

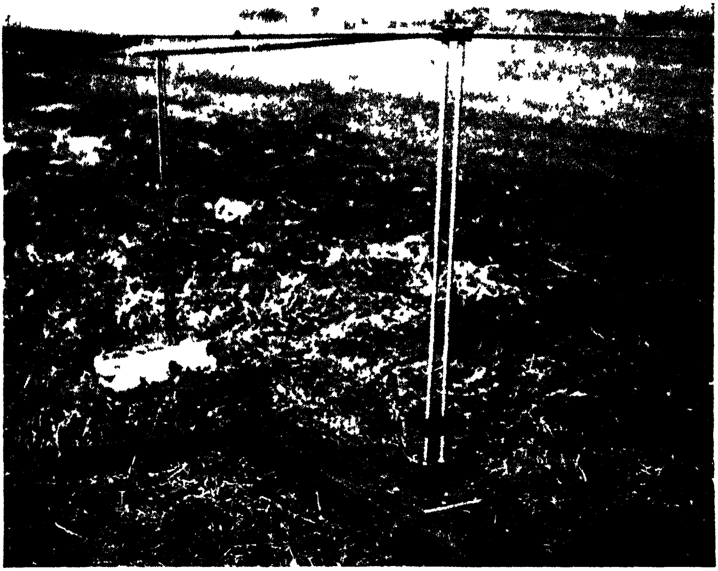


FIG. 9.—110 Mc half-Wave dipole at Central Radio Propagation Laboratory, National Bureau of Standards.

$$\overline{EN} = \overline{NF'} - \frac{\overline{NF}}{L_r} + 1 \tag{3}$$

We see by this relation that the two measurable quantities necessary to determine $\overline{EN}$ are $\overline{NF}$ and $\overline{NF'}$. This factor, $\overline{EN}$, in effect gives the ratio of the effective noise temperature of the radiation resistance of the antenna to the temperature of the signal generator load resistance taken as 300°K. The factor L_r enters into the expression as an allowance for the reduction in the signal energy caused by the transmission line and antenna losses; in other words $(\overline{NF}/L_r)$ is the receiver noise figure referred to the antenna terminals rather than to the receiver terminals.

Fig. 9 shows a photograph of the antenna for 110 Mc set up at the Sterling, Va., field station of the National Bureau of Standards where measurements are being made. Two conductor air dielectric transmission line with polystyrene spacers is used to couple the antenna to the receiver. The antenna is supported on a quarter-wave metal insulator as has been widely used in radar antennas. Calibration of the antenna and measurement of the effective noise figure $\overline{EN}'$ is accomplished by locating the calibrating diode right at the antenna terminals and using the radiation resistance of the antenna as the diode load resistance. Filament and plate potentials for the diode are brought to the diode with leads through the metal tubes of the quarter-wave insulator. Fig. 10 shows a close-up of a special experimental ultra high frequency diode for measuring the actual noise figure, $\overline{NF}/L_r$, while using a dummy resistor of the same impedance as the antenna for the diode load impedance.

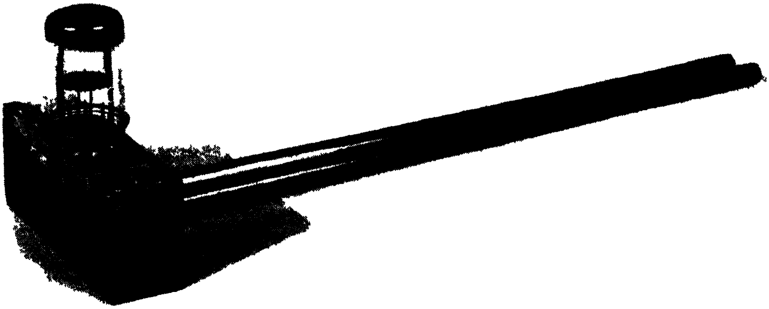


FIG. 10.—Very high frequency calibrating noise diode in place at antenna terminals.

Now, being in a position to measure the effective noise figure $\overline{NF}'$ and the actual noise figure $\overline{NF}/L_r$ with a dummy resistance at 300°K., we are able to determine the external noise factor, $\overline{EN}$. $\overline{EN}$ multiplied by 300 gives the effective absolute temperature of the radiation resistance of the antenna since, by definition, $\overline{EN}$ is simply the ratio of the available noise power of the antenna resistance to that of the dummy resistance at 300°. $\overline{EN} \times kTB$ $(T = 300°)$ gives the actual available received noise power at the antenna terminal. It should be pointed out that the antenna resistance could be at zero degrees absolute and N_a equal to zero if the antenna looked only at cold empty space, and could be many times 300° if the antenna were beamed on a very hot object emitting radiation on the frequency to which the antenna and receiver are tuned, e.g., Sagittarius at the frequencies on which cosmic radio noise has been received.

In all of these measurements, very sensitive receivers, that is, receivers with very low noise figures, are essential for appreciable deflection of a recorder in making measurements of cosmic noise. One of the low noise figure circuits for a radio frequency amplifier is shown in Fig. 11. This is known, after its inventor, as the Wallman low noise circuit. It consists of a conventional cathode separation amplifier tube loaded by the relatively low impedance input circuit of the following grounded grid stage. Neutralization of the first tube, while not absolutely essential, lends to the stability of the stage and decreases the loading of the grid circuit of the first tube. By proper adjustment of the input circuit, a noise figure of approximately 3 at 110 Mc has been obtained at the Bureau of Standards using a receiver employing two Wallman radio frequency amplifier stages preceding a triode converter with output on 5 Mc. Fig. 12 is a photograph of one of these converters for use in the cosmic radio noise program at the Bureau of Standards. At the right are the antenna connections followed from right to left by the two Wallman Stages, converter and local oscillator. Commercial high frequency

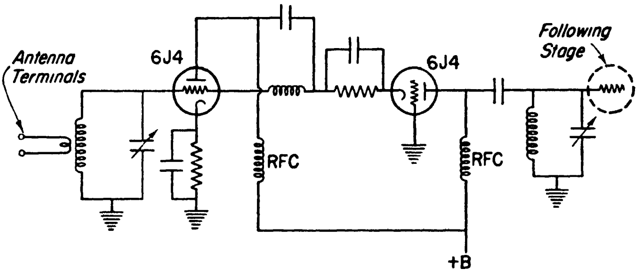


FIG. 11.—Neutralized Wallman low noise radio frequency amplifier.

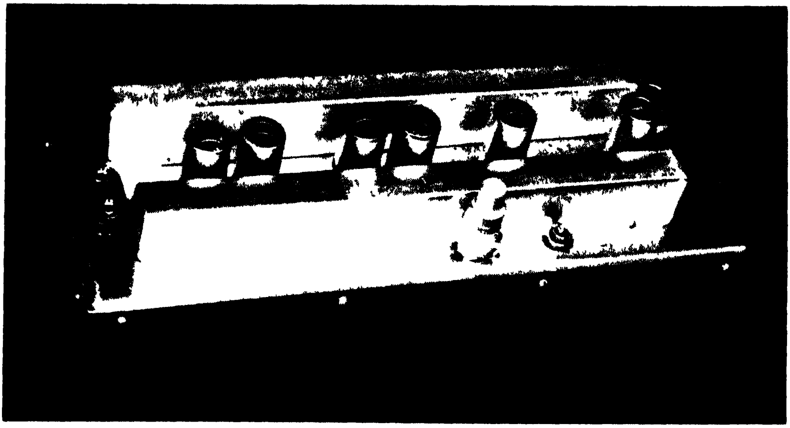


FIG. 12.—Very high frequency, low noise figure converter.

receivers with recording circuits installed provide the additional voltage gain necessary at an intermediate frequency to drive an Esterline Angus recorder.

VIII. RESULTS OF MEASUREMENTS

Results of measurements made by H. V. Cottony, W. Q. Crichlow, J. W. Herbstreit, and J. R. Johler at the Central Radio Propagation Laboratory are given in Fig. 13. This shows the received external radio noise power as a function of time on several days in August, 1947 for the frequencies 25 and 110 Mc. The external noise power is shown relative to kTB with the absolute temperature $T = 300°$. As the earth rotates, the cosmic radio noise increases and decreases in accordance with the part of outer space at which the antenna looks; the maximum cosmic radio noise occurs when the maximum response of the half-wave dipole is beamed in the general direction of the constellation Sagittarius. The solid curves shown are for the axis of the half-wave antenna oriented east-west so that the maximum response of the antenna beam is in the north-south plane. The dashed line shown for 110 Mc only is the external noise measured simultaneously on a half-wave antenna rotated 90° so that its maximum response is in the east-west plane. It may be seen that the solid lines for 25 and 110 Mc move up and down together and have the same general shape; however, the dashed line has a different characteristic shape, and a lesser maximum and higher minimum than for the east-west oriented antenna beamed more favorably on Sagittarius, the noisiest region in the Milky Way. Previous measurements made in June, 1947 with the same antenna as used to obtain the dashed curve shown are similar when compared at the same sidereal or star times.

Also shown in these measurements is the occurrence of several short time bursts of very strong solar radiation which were also accompanied by what are commonly referred to as sudden ionosphere disturbances or SID's.[27] These SID's manifest themselves on the high frequency bands as short period radio blackouts, on the daylight side of the earth, lasting a few minutes to possibly several hours. Two of these occurred during the measurements shown and both were accompanied by bursts of solar noise. The first occurred at 11:15 A.M. EST on August 23rd, a burst of solar noise occurring both on 25 and 110 Mc at the onset. The background level of cosmic radio noise did not change appreciably at 110 Mc following the burst of solar noise; however, at 25 Mc the background noise level dropped rapidly and recovered slowly indicating that a portion of the cosmic radio noise energy arriving at the earth was being absorbed at a low temperature before reaching the antenna. Jansky reported a similar occurrence while making his measurements in 1937.[3]

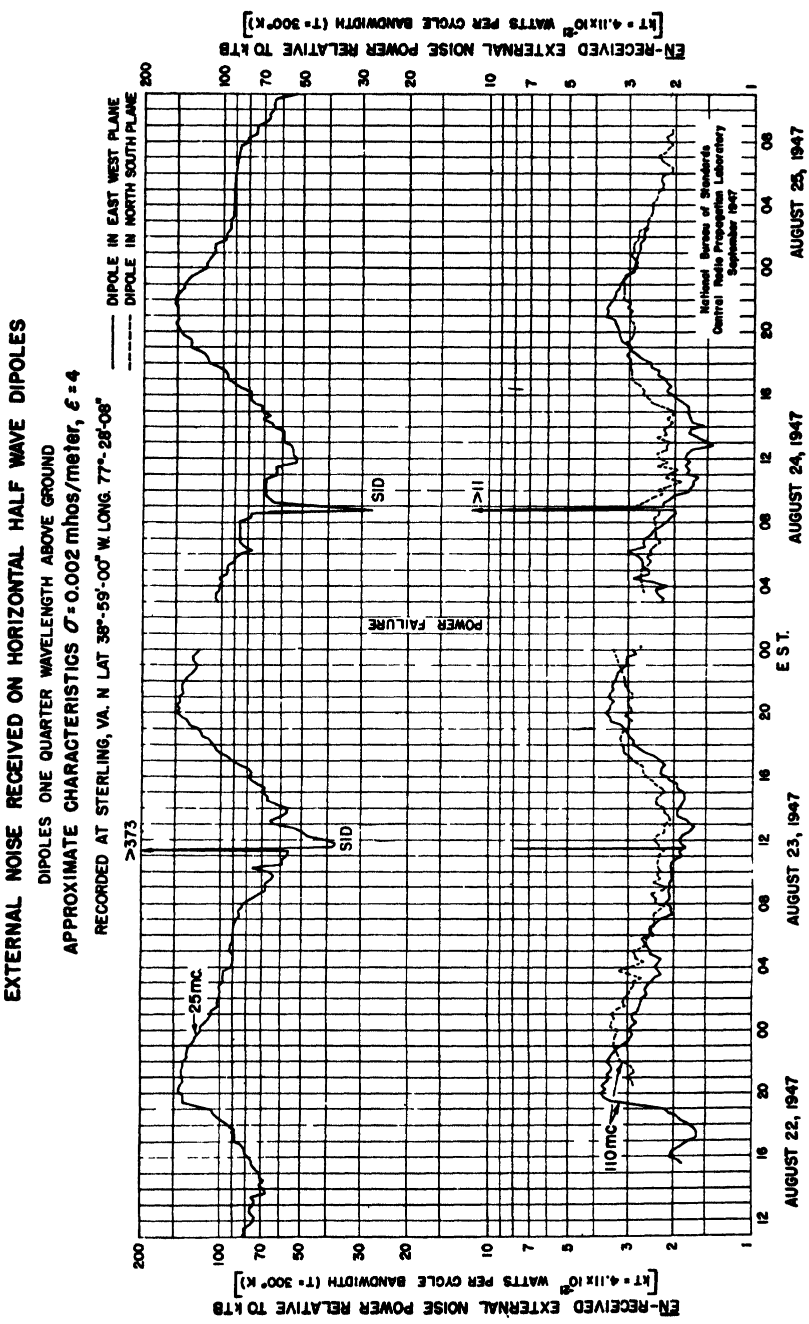


FIG. 13.—External noise received on horizontal half wave dipoles and recorded at Sterling, Va.

The radio blackouts occurring during these SID's are attributed to high ionospheric absorption which is undoubtedly the cause for the drop in external noise occurring only on 25 Mc, the absorption in passing through the ionosphere being, as would be expected, comparatively negligible at 110 Mc.

The second SID shown on this figure occurred at 8:30 A.M. August 24th and is somewhat different than the first in that the rapid drop in background external noise identified with high ionospheric absorption preceded the burst of solar noise by about 7½ minutes. Similar measurements made in England by Hey, Phillips, and Parsons have recently been reported.[14] The reason for the observed delays is not presently understood. One possible explanation suggested by photographs of eruptions on the sun is that an eruption starting near the sun's disc emits ionizing radiation at the start of the SID. Only the optical portion (possibly in the ultraviolet region) of the energy escapes through the densely ionized atmosphere of the sun. In a period of minutes, the eruption expands to the lesser ionized outer portions of the sun's atmosphere, which would permit the escape of lower frequency radiation and thus permit the delay in the arrival of noise energy observed at the lower radio frequencies. It is clear that studies of this type will be a very useful tool for investigating the connection between solar radiation and our ionosphere.

IX. Analysis in Terms of External Noise Factors

Fig. 14 shows the effective noise figures for receivers, including the effects of cosmic, ground, and receiver noise using horizontal half-wave receiving antennas one-quarter wavelength above the ground. The circled points at 25 and 110 Mc are the measured maximum and minimum values of external noise, $\overline{EN}$, at these frequencies. The received energy at the antenna is made up of direct and ground reflected waves, and since the ground is an imperfect reflector, a portion of the energy from the sky is absorbed in the ground in the form of heat and reradiated to the antenna at the temperature of the ground or 300°K. The resultant noise of the antenna is thus made up of two components, the incident energy from the sky and the portion radiated by the ground. The dashed line at the bottom of the figure gives the calculated noise radiated from the ground at a temperature of 300°. The ground external noise factor $\overline{EN}''$ is actually the portion of the total received energy that comes from the ground. The portion received from the sky at the actual sky temperature is equal to one minus the portion received from the ground. At frequencies less than approximately 200 Mc the effective temperature of the cosmic radio noise incident on the earth is in general greater than the temperature of the earth. When the energy from the two sources

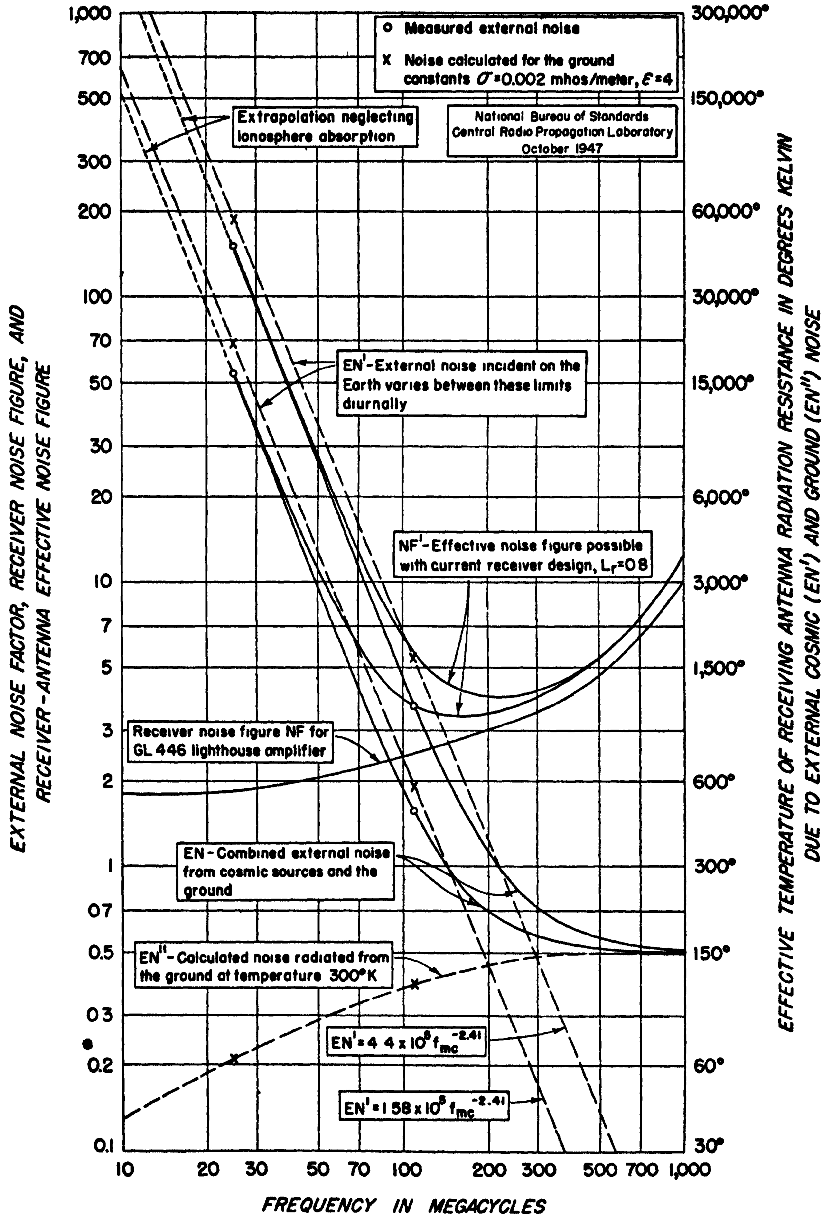


FIG. 14.—Effective noise figures for receivers including effects of cosmic, ground, and receiver noise using half-wave receiving antennas one-quarter wavelength above ground.

are added together in the proper proportion, the resultant external noise factor, $\overline{EN}$, lies along the solid lines through the circled points and merges into the ground radiation curve at the higher frequencies. The points denoted by X just above the circles at 25 and 110 Mc are the values of $\overline{EN}'$ calculated by removing the contribution of the ground from the measured values of $\overline{EN}$. The two dashed lines running diagonally down the figure represent an estimate of the way in which the maximum and minimum cosmic radio noise received from the sky varies with frequency, the slope of these curves being -2.41, the exponent in the assumed frequency law. As previously mentioned, the solid lines through the circled points, which merge into the ground radiation curve, give the combined external noise from cosmic sources and the ground as a function of frequency as received on a half-wave antenna one-quarter wave above an imperfect ground. These curves have been extrapolated down from 25 Mc to 10 Mc as indicated by the dashed portion of the curves neglecting the effects of ionosphere absorption, which in undoubtedly considerable at 10 Mc, at least in the daytime during the present period of very high sunspot activity.

Also shown in this figure is the variation with frequency of the receiver noise figure, $\overline{NF}$, for a GL446 lighthouse amplifier in a grounded grid circuit. The effective noise figure, $\overline{NF}'$, which results from the combination of all the factors, cosmic, ground, and this particular receiver noise, assuming a nominal transmission line loss, is also given for maximum and minimum cosmic radio noise conditions.

It is clear from the figure that at frequencies below approximately 100 Mc external cosmic radio noise is the prime factor which determines the minimum usable field intensity for a communication service, when using half-wave antennas a quarter wavelength above the ground with receivers of current design. Above approximately 100 Mc the current receiver sensitivity determines the minimum usable field intensity. The difference between $\overline{NF}'$ and $\overline{EN}$ gives the improvement possible with better receiver design as a function of frequency. Thus at 1000 Mc, it may be seen that an improvement of signal-to-noise ratio of approximately 24 times may be obtained with an ideal receiver where at 10 Mc improvement of receiver sensitivity will give no appreciable improvement. With very low noise figure receivers, further improvement may also be obtained at the higher frequencies with directional antennas which discriminate against the noise radiation from the ground. The use of directional antennas will provide an improvement in signal-to-noise ratio at the lower frequencies when receiving from directions from which little cosmic radio noise is arriving.

X. Field Intensities Required for Communication Services

In connection with the discussion of required field intensities for frequency modulation broadcasting given by Norton (see page 387) the effective noise figures shown in Fig. 14 have been translated into field intensities required for satisfactory FM broadcast reception; these required field intensities are shown as a function of frequency in Fig. 4, page 389. These curves may also be used to determine the required field intensity for other communication services by applying the appropriate bandwidth and signal-to-noise ratio factors as indicated in the figure.

XI. Effective Temperature Concept

If we consider the antenna radiation resistance to be in effective black body thermal equilibrium with the objects at which it is looking, and we know the gain of the antenna in all directions from which effective thermal radiation of temperature T is being received, then in accordance with this concept we may compute the effective temperature of the radiation resistance T_a as being equal to the mean temperature weighted in various directions in accordance with the antenna gain G. This relation is expressed mathematically as a surface integral thus:

$$T_a = 300\,\overline{EN} = \frac{1}{4\pi}\int\!\!\int_{4\pi} T(\theta,\phi)\cdot G(\theta,\phi)\,d\omega \tag{4}$$

where $T(\theta,\phi)$ is the absolute temperature of the material in space as properly weighted and averaged with respect to distance along the beam in an elementary solid angle $d\omega$ centered about the direction θ,ϕ. $G(\theta,\phi)$ is the gain of the antenna in the direction θ,ϕ. The proper method of determining the effective value of $T(\theta,\phi)$ may be seen most readily from the reciprocal problem in which energy radiated from the antenna is absorbed as it is propagated from the antenna out to a distance such that it is completely absorbed, and remembering that good absorbers are correspondingly good radiators. The following artificial example will serve to clarify the problem. If one-third of the total energy radiated in the elementary solid angle $d\omega$ centered on the direction θ,ϕ were absorbed in a gas of uniform absolute temperature of 300° extending from 0 to 1000 miles from the antenna, another one-third absorbed in a gas of uniform absolute temperature 30° extending from 1000 to 1250 miles, and the final one-third of the energy absorbed in a black body at the distance 1250 miles with a surface temperature of 600°, then $T(\theta,\phi)$ for that direction would be equal to $\frac{1}{3}(300 + 30 + 600) = 310°$ absolute. The definition of T_a as given by eq. (4), and as explained above, arises from the principle of detailed balancing in statistical mechanics, according to which in thermal equilibrium each infinitesimal process must be balanced by its inverse process.[28,29] Thus, since T_a has been confined by definition to refer only to temperature noise arising from fluctuations in the matter surrounding the antenna, and since the radiation resistance of the antenna must be in thermal equilibrium with its surroundings, it is only necessary to apply the principle of detailed balancing of absorption and radiation processes to the matter in each part of the space surrounding the antenna and to assume that $G(\theta,\phi)$ is the same for transmission and reception in order to derive eq. (4). It should be noted that, when $T(\theta,\phi)$ equals a constant value T_c in all directions, then T_a will simply be equal to T_c since the constant T_c may then be taken from under the integral signs and the integral is, by definition of $G(\theta,\phi)$, simply equal to 4π. On the other hand, for a high-gain antenna, if $T(\theta,\phi)$ has a very large value T_c over the effective beam of the antenna and a very small value in other directions, then T_a will again be nearly equal to T_c since the contributions to the integral for directions θ and ϕ far removed from the maximum of the antenna will be negligible. Measurements in the ultra-high-frequency band of the effective noise temperatures of antennas beamed on the open sky are of the order of 10° absolute, corresponding to the very low value of $\overline{EN} = 0.033$. When these antenna beams are directed horizontally along the ground, a small part, say one-tenth of the energy which could be transmitted from such an antenna would be absorbed in the ground, and, since the earth is at a temperature approximating 300°, the effective noise temperature of such an antenna used for reception and directed horizontally along the ground would be equal to $T_a = (\frac{1}{10})\,300 + (\frac{9}{10})10 = 39°$ corresponding to a value of $\overline{EN} = 0.13$. In the future, when receivers with very low noise figures become available in the ultra-high-frequency band, it may turn out to be desirable to discriminate against the ground-reflected wave in order to reduce the received noise.

XII. Distribution of the Intensity of the Noise Sources with Direction and Frequency

The use of the temperature concept outlined above implies that the radiation is thermal radiation following the well known black body radiation laws; whereas, actually the incident energy may be the resultant of a large number of discrete point sources of electromagnetic radiation not following black body radiation laws from stars such as our sun, or in interstellar space.

To better investigate this possibility and to evaluate results obtained with various antenna patterns located at various points on the earth, we will let $I(\theta,\phi)$ denote the intensity of the incident noise radiation from the direction θ,ϕ expressed in watts/sq. m./steradian as received in a 1 cycle band. The coordinate system has been chosen so that θ is the elevation angle and ϕ the azimuth angle with respect to the earth, as illustrated in Fig. 15. If we multiply the intensity $I(\theta,\phi)$ by the effective absorbing area, $A(\theta,\phi)$, and by the differential element of solid angle, $d\omega$, and then integrate over the entire antenna pattern, we obtain the total noise power, N_a, picked up by the antenna:

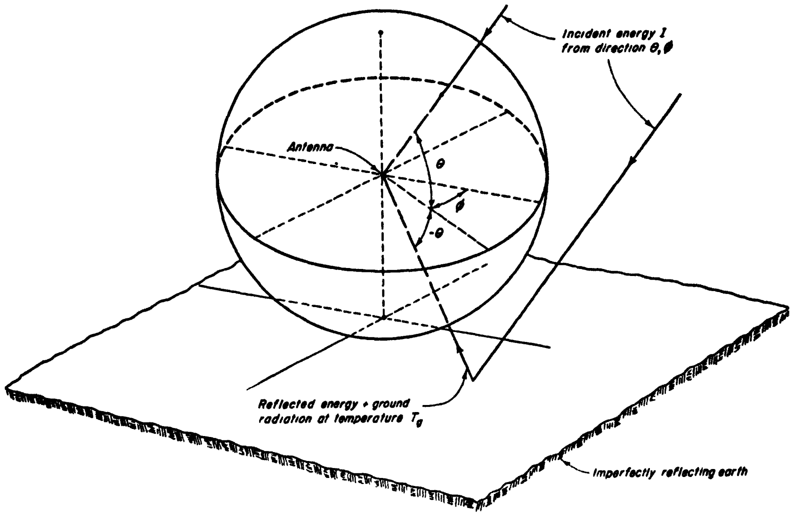

Incident energy I from direction θ,φ
Antenna
θ
φ
-θ
Reflected energy + ground radiation at temperature T_g
Imperfectly reflecting earth

FIG. 15.—Noise energy arriving at antenna from directions θ,ϕ and $-\theta,\phi$.

$$N_a = \int\int_{4\pi} I(\theta,\phi),\ A(\theta,\phi)d\omega \text{ watts in a 1 cycle band} \qquad (5)$$

It is to be understood in the above expression that the absorbing area, $A(\theta,\phi)$, is for a particular orientation of the antenna with respect to the earth and therefore N_a also corresponds to this particular antenna orientation. In other words, the noise power received with a directional antenna will depend upon the orientation of the antenna structure with respect to the earth. Furthermore, since cosmic radio noise originates in sources with fixed directions in the celestial sphere, the celestial coordinates of which change with time with respect to the point of observation on the earth, $I(\theta,\phi)$ will be a function of latitude and local time of the observer; consequently N_a will also vary with the latitude and local time of the observer.

From the above discussion it is seen that at least three independently varying coordinate systems are involved in the determination of N_a. The one is the coordinate system we have chosen, which is fixed with regard to the observers location on the earth's surface; a second is a system fixed with respect to the antenna pattern; and a third is the celestial coordinate system determining the locations of the cosmic noise sources. The evaluation of eq. (5) for a particular observer located at a particular point on the earth at a particular time, and employing an antenna oriented in a particular manner, can be made only by transforming the celestial coordinates for the noise sources into an expression for $I(\theta,\phi)$ and the antenna pattern coordinates into an expression for $A(\theta,\phi)$; so far only graphical methods have been used for the evaluation of the integral.

The effective absorbing area of any antenna may be expressed in terms of the gain of the antenna relative to an isotropic (omnidirectional) antenna by the following fundamental relation:

$$A(\theta,\phi) = \frac{G(\theta,\phi)\lambda^2}{4\pi} \text{ sq. m.} \tag{6}$$

In the above λ is the wavelength expressed in meters.

Substituting in the integral of eq. (5) we obtain the received noise power for a particular time and antenna orientation:

$$N_a = \frac{\lambda^2}{4\pi} \int\int_{4\pi} I(\theta,\phi)G(\theta,\varphi)d\omega \text{ watts in a 1 cycle band} \tag{7}$$

Since any receiving antenna has a mean absorbing area equal to $\lambda^2/4\pi$ when averaged over all directions in space, we may divide both sides of the above equation by $\lambda^2/4\pi$ and obtain an expression for the measured intensity of the incident radiation weighted in various directions in accordance with the antenna gain G. This measured incident radiation will be designated by M thus:

$$M \equiv \frac{4\pi N_a}{\lambda^2} = \int\int_{4\pi} I(\theta,\phi)G(\theta,\phi)d\omega \text{ watts/sq. m. for a 1 cycle band} \tag{8}$$

In particular it should be noted that in a hypothetical case in which it is assumed that the incident noise is constant in all directions, e.g., $I(\theta,\phi) = I_0$, this constant value may be removed from the integral and the resulting integral is then, by definition, simply equal to 4π so that:

$$M = \frac{4\pi N_a}{\lambda^2} = 4\pi I_0 \text{ (For noise arriving uniformly from all directions)} \tag{9}$$

In the general case, eq. (8) includes all of the noise energy arriving at the antenna from all directions, i.e., the cosmic noise radiation arriving from the hemisphere above the antenna, from the hemisphere below the antenna after being reflected from the surface of the earth, as well as the thermal radiation from the imperfectly reflecting earth also in the lower hemisphere. A physical representation of these components is shown in Fig. 15. To further illustrate this, we will divide eq. (8) explicitly into the three components representing: (1) the portion of the incident radiation arriving at the antenna directly from the direction θ,ϕ, (2) the portion of the incident radiation reflected from the ground and arriving at the antenna from the direction $-\theta,\phi$, this component must be combined with the direct radiation with the appropriate magnitude and phase as determined by the gain of the antenna in the direction $-\theta,\phi$, its height, h, above the ground and the ground reflection coefficient, $R(\theta,\phi)$, and finally (3) the portion of the thermal radiation received from the imperfectly reflecting ground at the temperature T_g. When eq. (8) is divided in this manner we have:

$$\begin{aligned} M = \int_0^{\frac{\pi}{2}} \int_0^{2\pi} \{I(\theta,\phi)|[&\overset{\text{(1) Direct radiation term}}{G^{\frac{1}{2}}(\theta,\phi)} \quad + \quad \overset{\text{(2) Ground-reflected radiation term}}{G^{\frac{1}{2}}(-\theta,\phi)R(\theta,\phi)e^{j4\pi h \sin\theta/\lambda}]^2|} \\ &\overset{\text{(3) Ground radiation term}}{+ G(-\theta,\phi)f(\theta,\phi)kT_gB/\lambda^2\}} \cos\theta d\theta d\phi \qquad (B = 1) \qquad (10) \end{aligned}$$

In the above $2h \sin \theta/\lambda$ is the path length difference, expressed in wavelengths, between the direct and ground-reflected waves and $f(\theta,\phi)$ is the fraction absorbed in the ground of the total radiation striking the ground from the direction θ,ϕ. The element of solid angle, $d\omega$, expressed in terms of the angles θ and ϕ, is equal to $\cos \theta d\theta d\phi$. The integration while taken only over a hemisphere, includes the total radiation arriving at the antenna from all directions. The portion of eq. (10) inside the absolute magnitude brackets is the expression for the directional pattern of the antenna for receiving cosmic noise; the solid models of antenna directivity shown in Fig. 6 were obtained by evaluating $|[G^{\frac{1}{2}}(\theta,\phi) + G^{\frac{1}{2}}(-\theta,\phi) R(\theta,\phi)e^{j4\pi h \sin \theta/\lambda}]^2|$ for half-wave dipoles one-quarter wavelength above (a) a perfect ground and (b) an imperfect ground. With a perfectly conducting ground, $R(\theta,\phi)$ is -1 and $f(\theta,\phi)$ is zero.

It is interesting to consider the case of a directional antenna having a maximum gain, G_m, which is taken as constant over the angular area, $\Delta\omega$, of its beam where, by definition, we will take:

$$\Delta\omega = \frac{4\pi}{G_m} \text{ steradians} \qquad (11)$$

In practice $\Delta\omega$ is approximately equal to the angular area of the beam at half-wave power response.[15] For these considerations, it is convenient to express the integral in eq. (8) as two terms, the first corresponding to integration over the major lobe of the beam, $\Delta\omega$, and the second being the integral over the remainder of the antenna pattern. Then M may be written:

$$M = \overline{I(\theta,\phi)}_{(\Delta\omega)} \int\!\!\int_{\Delta\omega} G(\theta,\phi)d\omega + \overline{I(\theta,\phi)}_{(4\pi-\Delta\omega)} \int\!\!\int_{4\pi-\Delta\omega} G(\theta,\phi)d\omega \qquad (12)$$

In the above equation $\overline{I(\theta,\phi)}_{(\Delta\omega)}$ and $\overline{I(\theta,\phi)}_{(4\pi-\Delta\omega)}$ denote appropriate weighted mean values of $I(\theta,\phi)$ averaged respectively over the main beam $\Delta\omega$ and over the remainder of the pattern $4\pi - \Delta\omega$. Now consider the integral

$$\int\!\!\int_{\Delta\omega} \overline{I(\theta,\phi)}_{\Delta\omega} G_m d\omega = I(\theta,\phi)_{\Delta\omega} G_m \Delta\omega \qquad (13)$$

where G_m is the maximum gain of the antenna. This integral may be added and subtracted from the right hand side of eq. (12) without changing its value and we obtain

$$M = \overline{I(\theta,\phi)}_{\Delta\omega} G_m \Delta\omega + \overline{I(\theta,\phi)}_{\Delta\omega} \int\!\!\int_{\Delta\omega} [G(\theta,\phi) - G_m] d\omega + \overline{I(\theta,\phi)}_{4\pi-\Delta\omega} \int\!\!\int_{4\pi-\Delta\omega} G(\theta,\phi)d\omega \qquad (14)$$

Since $G_m\Delta\omega = 4\pi$ by definition, we see that $\overline{I(\theta,\phi)}_{\Delta\omega}$ is numerically equal to $M/4\pi$ (when expressed as incident noise energy *per steradian*) if we neglect the two small correction terms involving the integrals; these are negligible, however, only for high gain antennas for which $\Delta\omega$ is small and we have in general:

$$\overline{I(\theta,\phi)}_{\Delta\omega} = \frac{M}{4\pi} - \frac{1}{4\pi} \overline{I(\theta,\phi)}_{(\Delta\omega)} \int\!\!\int_{\Delta\omega} [G(\theta,\phi) - G_m] d\omega - \frac{1}{4\pi} \overline{I(\theta,\phi)}_{(4\pi-\Delta\omega)} \int\!\!\int_{4\pi-\Delta\omega} G(\theta,\phi)d\omega \qquad (15)$$

In an effort to determine the frequency law of cosmic radio noise, there is shown in Fig. 16 the results of measurements made by several observers of the maximum received cosmic radio noise. In each case these maxima were observed when the maximum of the antenna beam was pointed approximately in the direction of the constellation Sagittarius and thus constitute an average of the noise from this direction plus that from other sources in the vicinity.

The circled values plotted are the measured values of the mean incident noise/steradian, $M/4\pi$, which were obtainable directly from the measured antenna noise power, N_a, received in a 1 cycle band simply by dividing by the wavelength, λ, squared. According to eq. (13), these values of M are also very nearly equal to the incident noise power, $\overline{I(\theta,\phi)}_{\Delta\omega}$, averaged over the area of the antenna beam $\Delta\omega$. In the case of the half-wave dipole measurements, however, the correction terms are not negligible; thus, for the National Bureau of Standards measurements appropriate corrections have been applied which yield the estimates of $\overline{I(\theta,\phi)}_{\Delta\omega}$ represented by the crossed points at 25 and 110 Mc. The solid line through these estimates of $\overline{I(\theta,\phi)}_{\Delta\omega}$ has a slope of -0.41 which is believed to represent the best presently available estimate of the expected variation with frequency for the intensity of the incident cosmic radio noise energy when expressed as a power law. It appears likely by Fig. 16 that, if Jansky's 1937 half-wave dipole measurements had been similarly corrected for ground absorption, they would have further substantiated this law. In the case of the other measurements, since the values of $\Delta\omega$ are different, an average over variable amounts of the sky is involved so that they would not be expected to yield the correct frequency law. However, since for these measurements the values of $\Delta\omega$ are all smaller than for the half-wave dipole measurements, they provide a more nearly correct estimate of the true maximum value of $I(\theta,\phi)$. The values of $\Delta\omega$ shown in the figure have been computed from eq. (11); the values of G_m were calculated from the physical dimensions of the arrays using the empirical relations given by Alford and Clarke.[15] G_m for the half-wave antenna over the ground was taken to be simply twice the maximum gain in free space, with no allowance for imperfect ground reflection.

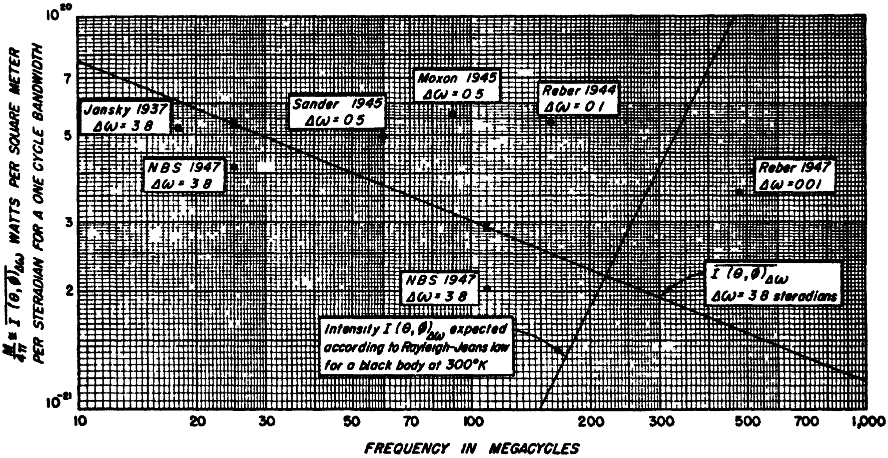


FIG. 16.—Maximum received cosmic radio noise/sq. wavelength/steradian.

XIII. INTENSITY FROM SMALL NOISE SOURCES

If $I(\theta,\phi)$ is negligible for all directions except those corresponding to the direction of a small source of area $\Delta\omega_s$, then for such a small source in the direction θ,ϕ we obtain by integrating eq. (8) over $\Delta\omega_s$

$$M = \overline{I(\theta,\phi)}_{\Delta\omega_s} \times G(\theta,\phi) \times \Delta\omega_s \text{ watts/sq. m. in a 1 cycle band } (\Delta\omega_s \leq \Delta\omega) \quad (16)$$

In terms of the effective temperature, T_s, of a black body source of area $\Delta\omega_s$, the incident noise power may be expressed, according to Rayleigh-Jean's law, as

$$\overline{I(\theta,\phi)}\Delta\omega_s = \frac{kT_sB}{\lambda^2} \quad (B = 1) \tag{17}$$

When the above is substituted in eq. (16) we obtain for the expected value of the measured incident noise:

$$M = \frac{G(\theta,\phi)\cdot\Delta\omega_s\cdot kT_sB}{\lambda^2} \text{ watts/sq. m. in a 1 cycle band} \tag{18}$$

XIV. OBSERVED INTENSITY OF RADIO FREQUENCY RADIATION FROM THE SUN

For some time considerable interest has been shown in measurements of radio frequency radiation from the sun. However, when the intensity of the radiation to be expected from the area of the sun ($\Delta\omega_s = 6.8 \times 10^{-5}$ steradians) in terms of black body radiation at its optical surface temperature of 6000°K. is considered as given by eq. (17), extremely high gain antennas are necessary before the power received from the sun at radio frequencies is detectable in the presence of receiver noise.

The first published measurements of solar radio frequency noise radiation were those of Reber at 160 Mc.[7] $\overline{I(\theta,\phi)}\Delta\omega_s$ determined from his results by using eq. (16) is approximately 350 times the expected black body radiation at this frequency for the sun at 6000°K. In making this determination the value of G was calculated from the actual physical area of Reber's antenna using the empirical relation given by Alford and Clarke[15]; $\Delta\omega_s$ was taken to be equal to the value observed visually at optical frequencies, i.e., 6.8×10^{-5} steradians.

Pawsey, Payne-Scott, and McCready in Australia have measured the quiescent average power radiated by the sun at 200 Mc to be approximately 170 times the expected value.[16] The quiescent effective intensity[17] of the sun is mentioned to distinguish it from large bursts of noise from the sun, which are associated with bright chromospheric eruptions and the general increase in solar noise which accompanies large groups of spots on the sun. During these bursts, enhanced radiation 6 million times the expected black body radiation has been observed at 45 Mc.[18]

Southworth at the Bell Telephone Laboratories has measured solar radiation in the frequency range from 3000 to 24000 Mc and has found the radiation to be approximately 5 times the expected 6000°K. value at 3000 and 10000 Mc, but only $\frac{8}{10}$ of this expected value at 24000 Mc, neglecting atmospheric absorption effects.[19] Dicke and Beringer reported measurements, including an atmospheric absorption correction, which correspond to 1⅔ times the expected value at 24000 Mc.[20] The results of these measurements of the quiescent radiation from the sun are shown in Fig. 17 together with the radiation to be expected from a black body with the visual area of the sun at an absolute temperature of 6000°K.

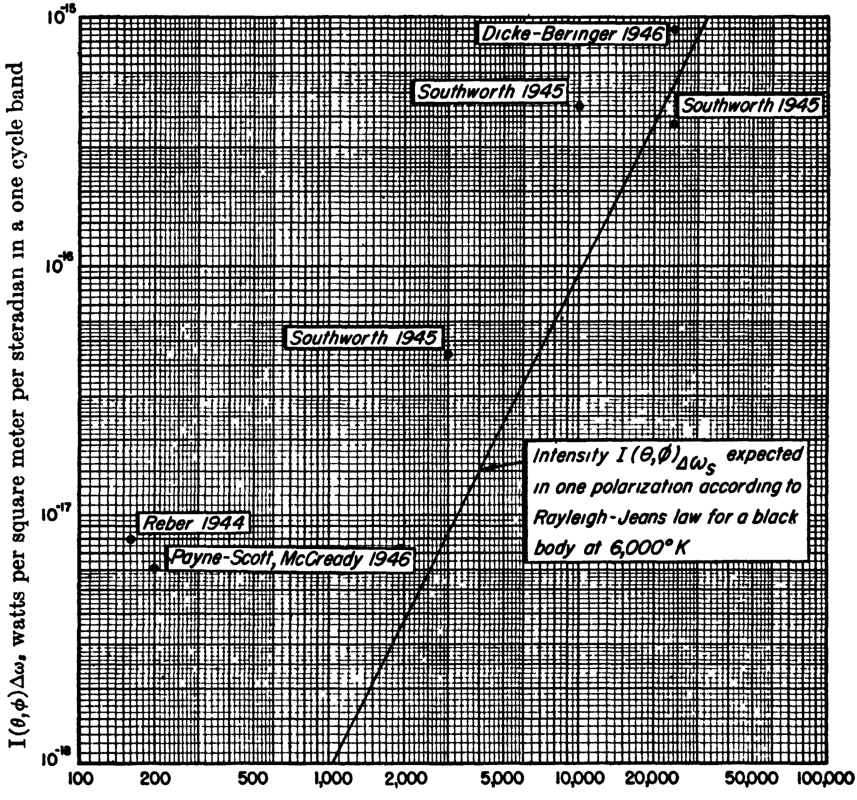


FIG. 17. —Measurements of radio frequency solar noise energy received in one polarization from the quiet sun.

To determine more accurately the source of the intense noise radiation from the sun at radio frequencies, the Australian experimenters have used, at 200 Mc, the method illustrated in Fig. 18.[21] This same method is being used to investigate the small source of cosmic radio noise in Cygnus. A directional antenna with a high gain was located on a cliff overlooking the sea to the east so that the interference pattern of the antenna caused by the combination of the direct and ground reflected waves was a series of lobes as shown. The antenna lobes were spaced by ½°, which happens also to be the angle subtended by the sun. Under these conditions, if the radiated noise energy were spread out over the entire surface of the sun, very little change in receiver output would be noticed after the sun rose and was passing through the lobe structure, since the measured integrated intensity, weighted in accordance with the antenna gain, would be approximately constant. However, when the sun rose, especially when a large sunspot was on the sun, the receiver output was not constant but a series of maxima and minima were observed corresponding to the antenna lobe pattern. In other words, when a spot was in a position corresponding to a maximum in the antenna lobe pattern, a maximum noise output was obtained and when it was in a null, a minimum was obtained. An analysis of the time of occurrence of these maxima and minima together with their amplitudes indicated that the noise was coming primarily from the direction of an individual sunspot or group of sunspots on the sun. In passing, it may be mentioned that information on the total refraction of these radio waves in passing through the earth's atmosphere has also been obtained from these measurements on the sun.

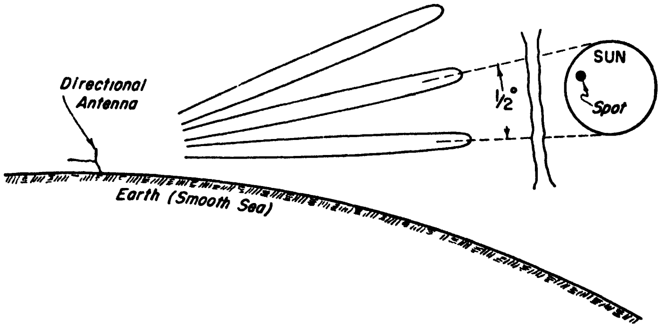


FIG. 18.—Australian method for resolving the source of sun noise.

XV. POLARIZATION OF EXTRATERRESTRIAL RADIATION

Observations of cosmic radio noise made up to the present time have not indicated that the incident noise energy has any systematic polarization (plane, circular, or elliptical) but that it is apparently randomly polarized. Consequently, observations have in general been made, and the preceding analysis has been given for the intensity of radiation incident upon an antenna sensitive to only one type of polarization. The form of the Rayleigh Jeans law given in eq. (17) also gives the intensity of the energy received in only one type of polarization. With randomly polarized waves, the total incident energy will be twice the energy received with only one type of polarization.

XVI. ORIGIN OF COSMIC RADIO NOISE

The possible origin of cosmic radio noise has been a subject of much speculation. Two main schools of thought exist on the subject today. One is that the incident cosmic radio noise energy results from intense noise radiation from eruptions on all of the stars in the universe similar to those associated with sunspot eruptions observed on the sun.[22] The other is that the noise originates in the extremely dilute matter occupying the very extensive space between the stars and arises from the random collisions of ionized electrons and matter existing there.

Greenstein, Henyey, and Keenan[23] have argued that the density of stars is not nearly great enough to account for the radiation in terms of the radiation associated with sunspot eruptions similar to those observed on the sun.

Townes[24] of Bell Telephone Laboratories has prepared the most recent review on this subject. He applies Kramers'[25] classical derivation for the continuous x-ray emission produced by bombarding nuclei with electrons to the assumed density and effective temperature of electrons in interstellar space to obtain the expected radiation. He points out that it is difficult to explain the observed radiation by this mechanism alone unless unusually high temperatures of the order of 100,000°K. are assumed to exist in interstellar space, having an electron density of 1/cc., whereas the generally accepted conditions of temperature are 10000°K. at a density of 1/cc. Townes also points out that Kramers' theory predicts that the intensity $I(\theta,\phi)$ will be independent of frequency at sufficiently high frequencies, and will vary as the square of the frequency at low frequencies. The National Bureau of Standards measurements, however, indicate that the intensity of the incident radiation decreases with frequency, at least in the 25 to 110 Mc range.

It is probable that both mechanisms mentioned above and possibly some others not as yet considered are responsible for the observed cosmic radio noise and at this point, the reader is referred to another recent summary[26] and his own imagination.

The fact that cosmic and solar radio noise is arriving at the earth with sufficient intensity to determine the lowest usable field intensities for radio communication services throughout a wide band of frequencies is now well established. It is extremely difficult to say just what other direct effects this relatively newly discovered phenomena may have on ourselves. However, it is clear that an entirely new approach to the workings of the sun and the universe is open for investigation through the study of cosmic and solar noise.

REFERENCES

1. Jansky, K. G. *Proc. Inst. Radio Engrs.*, **20,** 1920 (1932).
2. Jansky, K. G. *Proc. Inst. Radio Engrs.*, **21,** 1387 (1933).
3. Jansky, K. G. *Proc. Inst. Radio Engrs.*, **25,** 1517–1530 (1937).
4. Reber, G. *Electronic Inds.*, **3,** 89–92 (1944).
5. Reber, G. *Proc. Inst. Radio Engrs.*, **30,** 367–378 (1942).
6. Friis and Feldman. *Proc. Inst. Radio Engrs.*, **25,** 841 (1937).
7. Reber, G. *Astrophys. J.*, **100,** 279–287 (1944).
8. Hey, J. S., Phillips, J. W., and Parsons, S. J. *Nature, Lond.*, **157,** 296 (1946).
9. Sander, K. F. Measurement of Cosmic Noise at 60 Mc/s, Radar Research Development Establishment Report No. 285, May 1945.
10. Moxon, L. A. Galactic Noise Measurements on 90 Mc/s, Admiralty Signal Establishment Extension Report, Reference XRC 3/45/9, dated 26/11/1945.
11. Hey, J. S., Parsons, S. J., and Phillips, J. W. *Nature, Lond.*, **158,** 234, (1946).
12. Friis, H. T. *Proc. Inst. Radio Engrs.*, **32,** 419–422 (1944).
13. Nyquist, H. *Phys. Rev.*, **32,** 110–113 (1928).
14. Hey, J. S., Parsons, S. J., and Phillips, J. W. *Nature, Lond.*, **160,** 371 (1947).
15. Alford, A., and Clarke, I. G. Semi-empirial Relations Between the Gain, Aperture, Beam Width and Shape of High Gain Antennas, Harvard University, Radio Research Laboratory Report 411–119, October 19, 1944.
16. Pawsey, J. L. *Nature, Lond.*, **158,** 633 (1946).
17. Martyn, D. F. *Nature, Lond.*, **158,** 632 (1946).
18. Norton, K. A., and Omberg, A. C. *Proc. Inst. Radio Engrs.*, **35,** 4–24 (1947).
19. Southworth, G. C. *J. Franklin Inst.*, **239,** 285 (1945).
20. Dicke, R. H., and Beringer, R. *Astrophys. J.*, **103,** 375 (1946).
21. McCready, L. L., Pawsey, J. L., and Payne-Scott, R. *Proc. Roy. Soc.*, **190,** 357–375 (August 12, 1947).
22. Pawsey, J. L., Payne-Scott, R., and McCready, L. L. *Nature, Lond.*, **157,** 158, (1946).
23. Greenstein, J. L., Henyey, L. G., and Kennan, P. C. *Nature, Lond.*, **157,** 805 (1946).
24. Townes, C. H. *Astrophys. J.*, **105,** 235 (1947).
25. Kramers, H. A. *Phil. Mag.*, **46,** 836–871 (1923).
26. Reber, G., and Greenstein, J. L. *Observatory*, **76,** 15 (1947).
27. Ionospheric Data, CRPL-F37, September 1947, a monthly report issued by the Central Radio Propagation Laboratory of the National Bureau of Standards, Washington, D. C.
28. Slater, J. C. Microwave Transmission, McGraw-Hill, New York and London, 1942, p. 256.
29. Slater, J. C. Introduction to Chemical Physics, McGraw-Hill, New York and London, 1939, p. 91.

Propagation in the FM Broadcast Band

KENNETH A. NORTON

Central Radio Propagation Laboratory,
National Bureau of Standards,
Washington, D. C.

CONTENTS

		Page
I.	Introduction	381
II.	The Interference Due to Long Distance Ionospheric Propagation	382
III	The Effects of Radio Noise on Broadcast Reception	387
IV	The Effects of Antenna Height and Terrain on the Effective Transmission Range Over a Smooth Spherical Earth	390
V	The Effects of Irregularities in the Terrain	391
VI.	The Systematic Effects of Terrain and of Tropospheric Ducts	392
VII	The Tropospheric Waves Resulting from Reflection at Atmospheric Boundary Layers	402
VIII.	The Combined Effects of Ducts and of Random Tropospheric Waves	408
IX.	The Calculated Service and Interference Ranges of FM Broadcast Stations	410
X.	The Efficient Allocation of Facilities to FM Broadcast Stations	413
XI.	The Optimum Frequency for an FM Broadcast Service	414
	References	421

I. INTRODUCTION

Frequency modulation broadcasting is now a reality in dozens of communities throughout the United States. Within another year, the number of FM broadcasting stations actually operating will probably exceed the number of AM broadcasting stations which were operating before the war. The experience so far with this new broadcasting service has been very favorable. FM was expected to be superior to AM for broadcasting in two major respects: a greater potential audio frequency range of transmission and a greater potential volume range of transmission due to a greater freedom from radio noise and interference from other radio stations. The potentially greater audio frequency range has been realized to a great extent already, and many additional improvements in this direction may be expected in the near future in view of the extensive work now in progress on audio frequency amplifiers and loud speakers. The potentially greater volume range of transmission with FM is due largely to the differences in the propagation in the FM and in the standard

AM broadcast bands. These propagation factors are the subject of this paper.

In the first place, it is important to emphasize that it is not enough for FM to be a slightly better broadcasting service. It must provide a very greatly improved service. Otherwise, the public can scarcely be expected to support it. For this reason, the Federal Communications Commission has in the past taken the view that the standards of good engineering practice must be very much higher for this new service. In particular, interference from other AM broadcasting stations is considered serious if it exists for more than 10% of the time while the corresponding tolerance with FM is only 1%. Using this higher standard, a discussion will be given of some of the things now known about FM broadcast propagation and mention made of some of the unsolved problems to the solution of which future research might well be directed.

II. THE INTERFERENCE DUE TO LONG DISTANCE IONOSPHERIC PROPAGATION

It is a popular belief that propagation on the frequencies used for FM broadcasting is limited in range approximately to line-of-sight distances, and consequently to much shorter distances than in the AM broadcast band where the propagation is characterized by surface waves which follow the curved surface of the earth to distances well beyond the line-of-sight. This is true to the extent that high powered stations on clear channels are available for AM broadcasting but, in practice, several stations are permitted to operate simultaneously on most of the AM channels. The result is that the effective range of each such station is drastically reduced by mutual interference between the several stations occurring at night when conditions for long distance ionospheric transmission are favorable. A good example of this difference in propagation in the AM and FM broadcast bands is available at the author's home in Washington, D. C. Radio Broadcast Station WINC in Winchester, Va., a distance of about 50 miles from Washington, broadcasts the same programs on AM and FM. The FM broadcasts are satisfactory both day and night but the AM broadcasts are usable only in the daytime and frequently suffer such intolerable interference at night that the program cannot even be identified. Thus, the FM station provides service for more than 99% of the time at my home while the AM station fails to provide even a 90% service at distances greater than a few miles. Thus, it turns out that stations operating in the FM broadcast band should ordinarily be able to cover much larger areas than stations operating in the AM broadcast band, except in those cases where the AM stations enjoy the privilege of operating on a clear channel. The reason for the

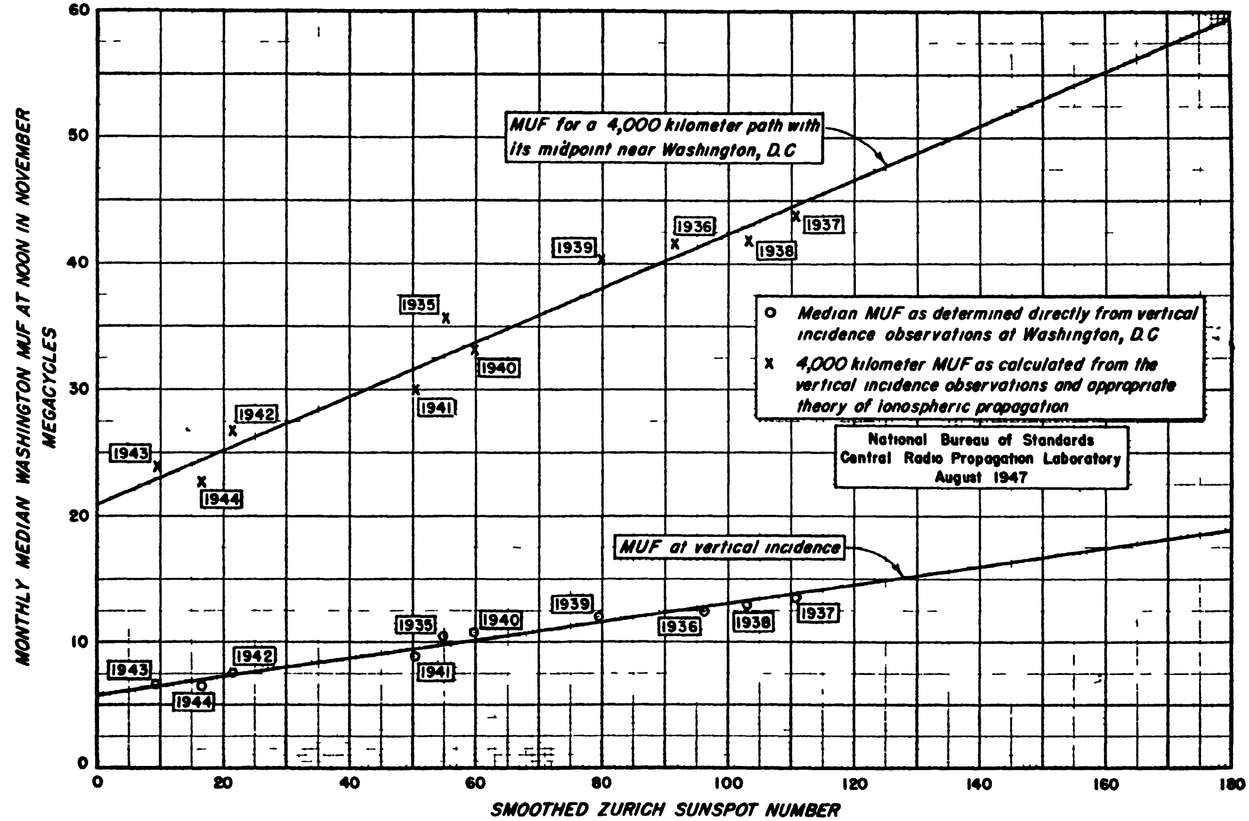


FIG. 1.—Correlation of the median Washington noon maximum usable frequency for November with the smoothed sunspot number. (Courtesy of *Q.S.T.* Magazine.)

superiority of the FM frequency band in this case is the fact that the ionosphere will not support transmission on the VHF frequencies 88–108 Mc now in use for FM broadcasting. When the frequency of transmitted radio waves is increased continuously, a maximum frequency is finally reached above which the ionosphere no longer reflects the waves back to the earth; this maximum frequency is known as the maximum usable frequency and it is known to depend upon many factors, the principal ones being the distance between the transmitter and receiver, the geographic and geomagnetic latitudes, the season of the year and local hour for the midpoint of the transmission path, and the solar activity. For transmission paths within the United States, the highest frequencies are reflected from the F layer of the ionosphere near the middle of the day in November at latitudes between 35 and 40°. Fig. 1 shows the very close correlation between these maximum usable frequencies and the solar activity as measured by the smoothed sunspot numbers published by the Zurich Observatory in Switzerland.[1] The points plotted are the values of the monthly median maximum usable frequencies for November as determined by ionospheric measurements made near Washington, D. C. The separate points correspond to measurements made in successive years from 1935 to 1944 during which the sunspot activity varied on the Zurich scale from 10 to 110. The upper curve is for transmission over a 2500-mile path while the lower curve corresponds to propagation at vertical incidence.

Fig. 2 shows the variations of sunspot activity as determined by the Zurich astronomers for the past 200 years. This figure shows the regular cycle of solar activity with a period of approximately 11 years, together with the longer trend exhibited by a 49-year running average which smooths out the effect of the 11-year cycle. Fig. 3 shows the most recent variations in the monthly mean and the smoothed sunspot numbers; the smoothed sunspot numbers are moving averages for 13 successive months and are shown by the solid line. Also shown on Fig. 3 are several values of the smoothed sunspot number as predicted for the period near the coming maximum of sunspot activity; the McNish and Lincoln[2] predictions were made by the statistical method in current use at the Central Radio Propagation Laboratory of the National Bureau of Standards in connection with the regular forecasts of high-frequency transmission conditions;[3] the Waldmeier[4] prediction was made late in 1945; and the Stewart[5] prediction was communicated to the Central Radio Propagation Laboratory in a note dated June 14, 1946. The points indicated by the triangles on Fig. 3 are known as ionospheric sunspot numbers and represent a measure of the sunspot activity obtained from the charac-

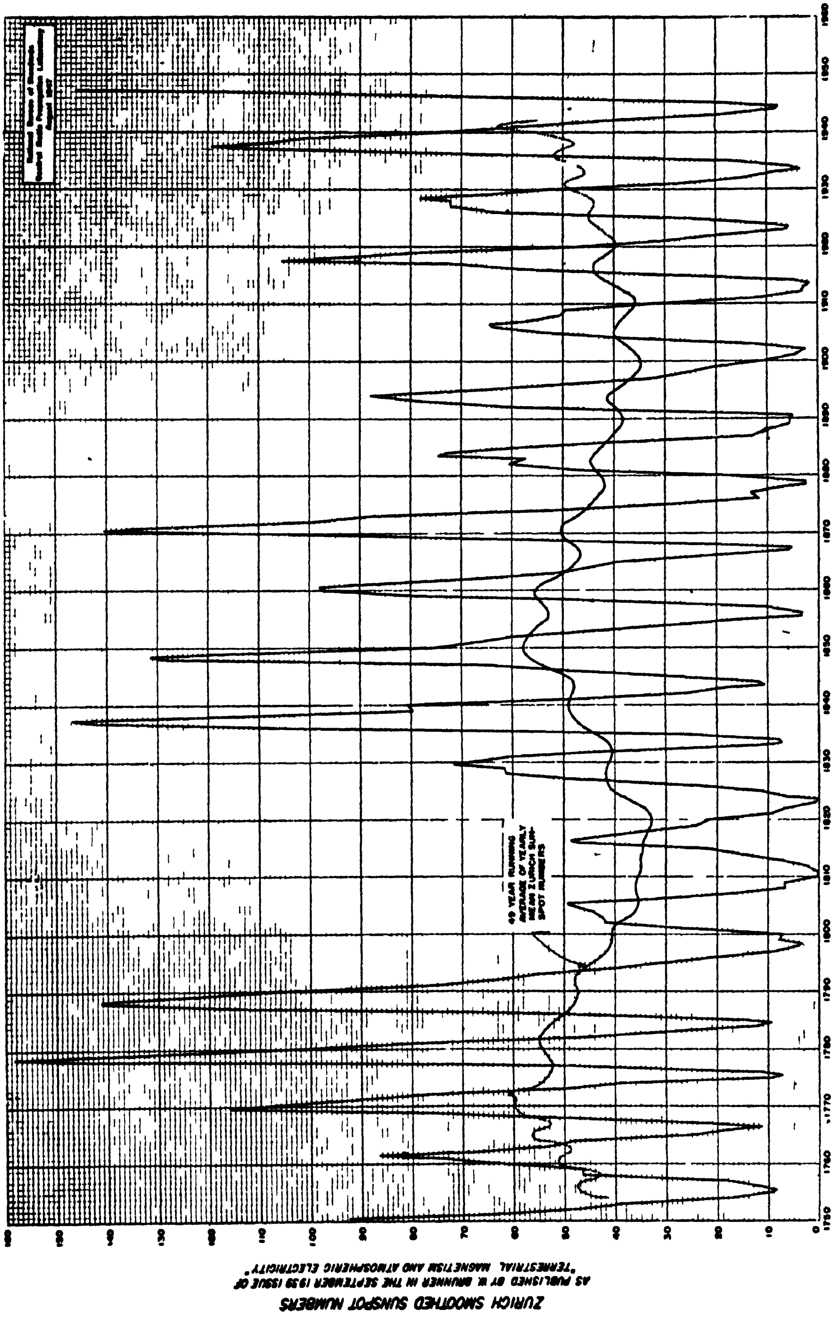


FIG. 2.—Secular variations exhibited by past sunspot cycles. (Courtesy of *Q.S.T.* Magazine.)

teristics of the ionosphere as observed at various ionospheric monitoring stations throughout the world.[6]

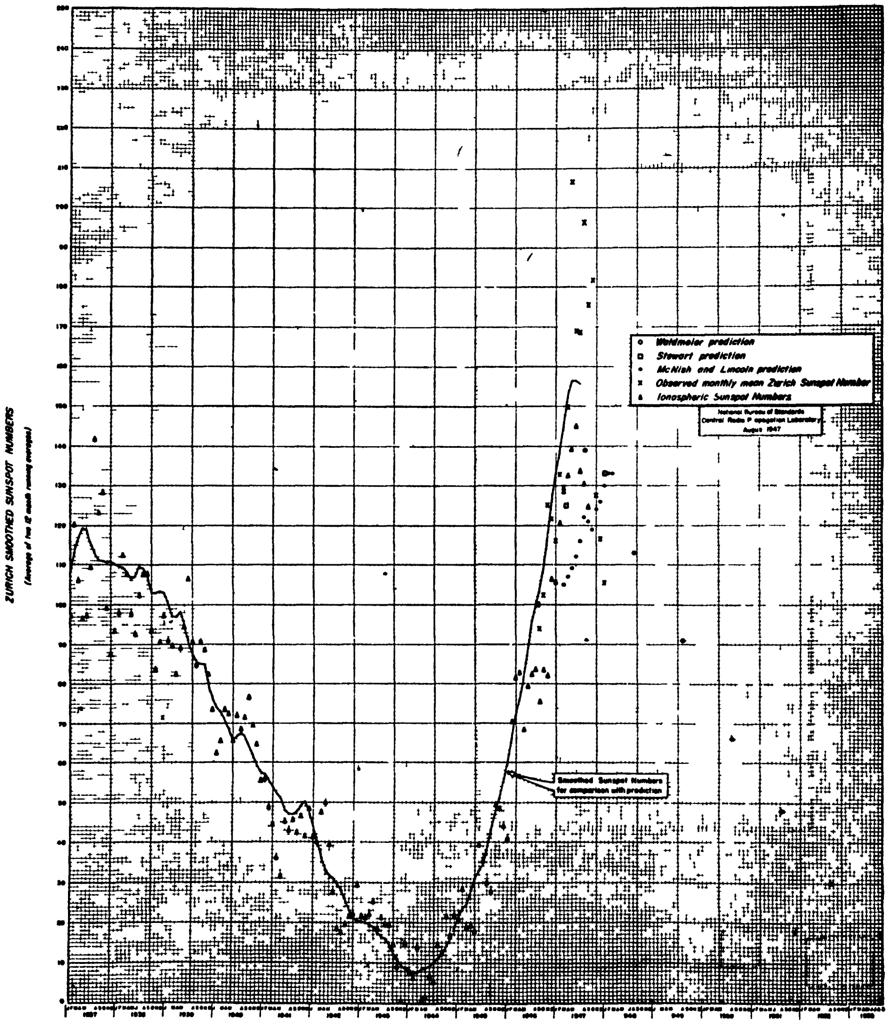


FIG. 3.—Recent values of sunspot activity with predictions for coming months. (Courtesy of *Q.S.T.* Magazine.)

Having established the approximately linear relation between the maximum usable frequencies and the smoothed sunspot numbers shown on Fig. 1 it is possible to estimate the maximum usable frequencies which may be expected for other values of sunspot activity. In particular, since the highest sunspot number ever observed in the past was less than

160 we see that median values of the maximum usable frequencies are not likely to exceed 55 Mc at Washington for any November in the future. It is interesting to note that the monthly sunspot activity for November 1947 was considerably higher than for any November since the invention of radio and consequently, it would be expected that the amateur 50 Mc band would be useful for F layer East-West Coast contacts for the first time; this was, in fact, found to be the case.[7] Further details regarding these ionospheric phenomena are given in the forecasts of ionospheric propagation contained in the monthly publication of the Central Radio Propagation Laboratory of the National Bureau of Standards entitled "Basic Radio Propagation Predictions."[8]

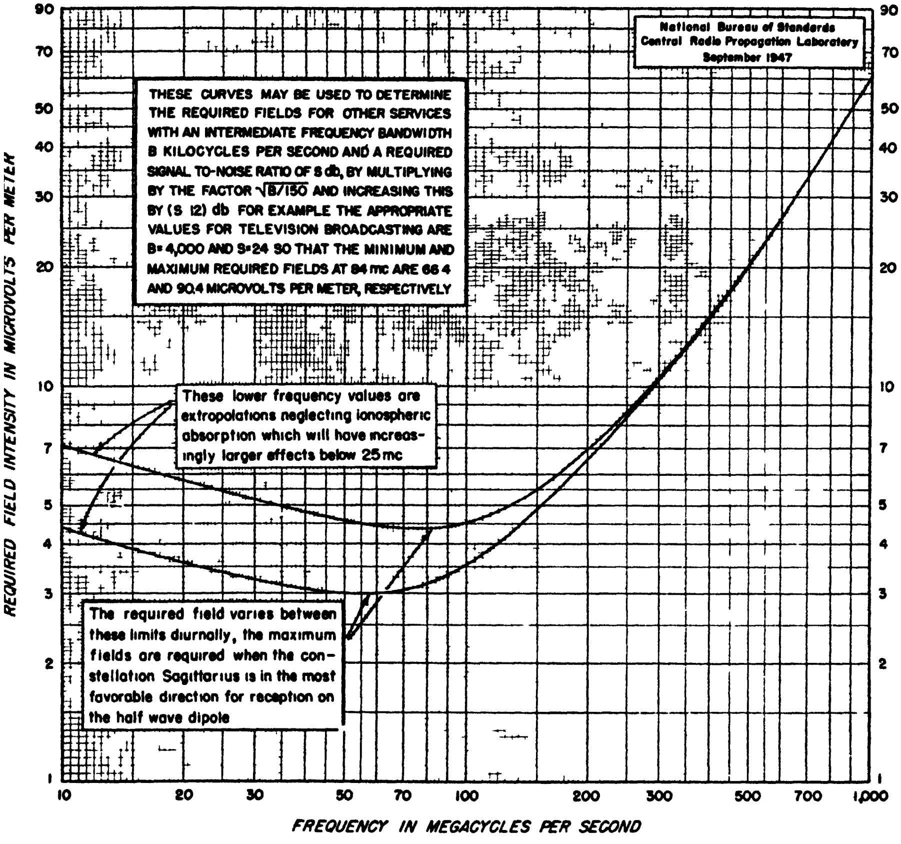
In addition to the regular F layer ionospheric transmission just discussed, sporadic ionospheric transmission may also take place by reflection from local concentrations of ionization which occur from time to time in the E region of the ionosphere. These latter transmissions are, in general, weaker and less predictable than the F layer transmissions, but are of equal importance, perhaps, as a source of interference to broadcasting because they occur every year and not just for the years near the sunspot maximum.

Fortunately, the Federal Communications Commission recently moved the FM broadcast band from below 50 Mc to its present position in the 88–108 Mc band and thus practically eliminated the possibility of both F layer and sporadic E layer interference. A very few scattered reports of long distance transmission in the 88–108 Mc band have been received and these are usually considered to be due to sporadic E layer reflections. However, experience to date indicates that ionospheric transmission may be expected in the 88–108 Mc band for very much less than the 1% of the time which has been adopted as a measure of the permissible interference to an FM broadcasting service. This absence of long distance interference is the principal difference in the propagation characteristics of FM and AM broadcasting frequencies, which makes possible much larger potential service areas with FM than with AM broadcasting. Unfortunately, these tremendous potentialities of FM broadcasting are not now being fully realized because of the necessity, in many rural and suburban receiving locations, for an adequate receiving antenna and a very sensitive receiver.

III. The Effects of Radio Noise on Broadcast Reception

The second major difference in the propagation characteristics of AM and FM broadcasts is the difference in the radio noise levels encountered in the two frequency bands. There are two cases to consider. The first is man-made radio noise which is usually strongest in the built-up city areas. Fortunately, the radio fields are usually also strongest in these same areas and this tends to overcome this type of interference. In this case, FM has the advantage over AM that the man-made radio noise intensity is usually weaker at these VHF frequencies. The worst man-made noise sources at FM frequencies are the ignition systems of automobiles, and it is encouraging to note that the auto manufacturers are at the present time giving serious consideration to the addition of noise suppressors as standard equipment. Such a step is obviously not only of great importance to the public generally, but it should also make the introduction of FM in auto radios a more successful venture. However, the greatest advantage of FM over AM as regards noise interference is found in suburban and rural areas where man-made noise is absent for a large part of the time. In these areas the noise that inherently limits reception is simply that due to natural sources such as atmospheric radio noise from thunderstorms in the troposphere and the cosmic radio noise originating in the interactions of matter in interstellar space together with direct radio noise radiation from the stars. The field intensities of atmospheric noise near their sources in the lightning flashes are known to vary inversely with the radio frequency and, as received, this atmospheric noise decreases with increasing frequency even more rapidly than this because of the more rapid rate of attenuation at the higher frequencies involved in propagation from the source to the receiver. Thus, in the FM frequency band atmospheric noise is only a factor during the very small percentage of the time that the thunderstorms occur in the immediate vicinity of the receiver, whereas in the AM frequency band atmospheric noise is the factor limiting the range of AM stations for nearly 100% of the time when a good receiving antenna and sensitive receiver are used for reception. It is this prevalence of atmospheric noise at AM frequencies which has led to the common use of small receiving antennas built in to the receiver for this frequency band; experience has shown that little improvement in reception may be expected with the use of a more elaborate antenna since such an antenna collects additional atmospheric noise in direct proportion to the additional signal, with little improvement in signal-to-noise ratio. In the VHF band, on the other hand, the limiting range of reception is usually determined by radio noise generated in the high-frequency circuits of the receiver itself rather than from external sources so that an improvement in the receiving antenna results in an almost proportional improvement in output signal-to-noise ratio.

In an effort to determine quantitatively just how important external noise sources are in the VHF and higher frequency bands, the Central Radio Propagation Laboratory of the National Bureau of Standards has initiated an extensive program of research along these lines. Some of the results of this work are presented by J. W. Herbstreit in this book; the results of particular interest in connection with FM reception are the cosmic noise measurements shown in Figs. 13 and 14, pages 365 and 367



THESE CURVES MAY BE USED TO DETERMINE THE REQUIRED FIELDS FOR OTHER SERVICES WITH AN INTERMEDIATE FREQUENCY BANDWIDTH B KILOCYCLES PER SECOND AND A REQUIRED SIGNAL TO-NOISE RATIO OF S db, BY MULTIPLYING BY THE FACTOR $\sqrt{B/150}$ AND INCREASING THIS BY (S 12) db FOR EXAMPLE THE APPROPRIATE VALUES FOR TELEVISION BROADCASTING ARE B=4,000 AND S=24 SO THAT THE MINIMUM AND MAXIMUM REQUIRED FIELDS AT 84 mc ARE 66 4 AND 90.4 MICROVOLTS PER METER, RESPECTIVELY

These lower frequency values are extropolations neglecting ionospheric absorption which will have increasingly larger effects below 25 mc

The required field varies between these limits diurnally, the maximum fields are required when the constellation Sagittarius is in the most favorable direction for reception on the half wave dipole

FIG. 4.—Field intensities required for satisfactory FM broadcast reception in the absence of thunderstorms or man-made noise Assumes use of receiver with a noise figure as in Fig 3, "Maximum Range of a Radar Set," *Proc Inst Radio Engrs*, January 1947, transmission line loss, $L_r = 0\ 8$, horizontal half wave dipole receiving antenna; and an intermediate frequency RMS signal to noise ratio of 12 db, the external noise is assumed to be cosmic noise varying in inverse proportion to the 2 4 power of the frequency as observed at Sterling, Va , intermediate frequency bandwidth = 150 kc.

From the data given on those figures it is possible to determine the field intensities which would be required for satisfactory FM broadcast reception in the absence of thunderstorms or man-made noise, and these are shown on Fig. 4. The curves are based on the assumption that a half-wave dipole receiving antenna is connected to a very good low-noise-figure receiver by means of a transmission line with a loss of 1 db. Two curves are given corresponding to the minimum and maximum recorded noise levels shown on Fig. 13, page 365. Fig. 4 indicates that 4 microvolts/meter will provide a satisfactory FM signal for this receiver, which was assumed to have a noise figure of only 2.4 at 98 Mc. Measurements on a small sample of FM receivers currently being offered to the public indicates that corresponding figures of required fields range from about 8 to 14 microvolts/meter; we see from this comparison that substantial further improvements in receivers are both possible and desirable. Based on this discussion, it will be assumed in the remainder of this paper that 10 microvolts/meter is a good average figure to use for the field intensity required in quiet rural areas to provide a satisfactory FM service; the use of this figure implies that a fairly good receiver is used in conjunction with an outside half-wave antenna.

IV. THE EFFECTS OF ANTENNA HEIGHT AND TERRAIN ON THE EFFECTIVE TRANSMISSION RANGE OVER A SMOOTH SPHERICAL EARTH

The next topic to consider is the prediction of the distance from the FM transmitting antenna at which the minimum required field may be expected. This problem is truly a formidable one involving, as it does, the transmitting antenna height and gain, the nature of the intervening terrain, the frequency on which the transmissions take place, the distribution with height of the refractive index of the air in the troposphere through which the radio waves must travel, and the height of the receiving antenna. Here again the nature of the propagation of the ground wave in the AM and FM broadcast bands is entirely different. In the AM broadcast band the transmitting and receiving antennas are necessarily at heights less than a wavelength above the ground and the received fields are surface waves.[8] The intensity of these surface waves is determined by the well-known Sommerfeld formula which shows that, at a fixed distance, the field intensity decreases in inverse proportion to the square of the frequency. On the other hand, in the VHF band, the transmitting and receiving antennas are usually elevated several wavelengths above the ground and the received fields are space waves.[8] Up to distances somewhat greater than line-of-sight these received VHF space wave fields actually increase in intensity with an increase in frequency. Thus, up to points slightly beyond the line of sight, the higher VHF frequencies tend to provide larger service areas. However, these conclusions are based on propagation over a smooth surface. Irregularities in terrain, including the bulge of the earth itself, have the effect of cancelling this advantage of the higher VHF frequencies since the expected fields far beyond the line of sight decrease with increasing frequency. As a good first approximation to this problem of estimating FM coverage we can calculate the field intensities to be expected with the given transmitting and receiving antenna heights over a smooth spherical earth surrounded by an atmosphere with a constant lapse rate of refractive index of 12×10^{-6} parts/thousand feet, corresponding approximately in field intensity calculations to the use of an effective earth's radius four-third's of its actual value. These smooth earth, standard atmosphere fields may be determined by the methods given in a recent paper by Norton;[8] for the FM broadcast band 88–108 Mc they are also included in the Standards of Good Engineering Practice of the Federal Communications Commission. These smooth earth, standard atmosphere fields must be corrected in individual cases to allow for irregularities in the terrain and the atmosphere which will always be present. The remainder of this paper deals with these variations.

V. THE EFFECTS OF IRREGULARITIES IN THE TERRAIN

FM transmitting antennas are frequently erected in the built-up business areas of the city, and unless these FM antennas are higher than the buildings around them, the field intensities may be expected to be considerably below the values calculated by using the height of the antenna above the local terrain and the smooth earth theory. These differences are greatest at the shorter distances. On the other hand, if the antenna is substantially higher than the buildings in its vicinity, then the observed ground wave field intensities are ordinarily in good agreement with the predicted smooth earth values, especially at great distances, unless the terrain near the receiving antenna is very hilly or the intervening terrain is markedly different from the gradual curvature of the earth assumed in the theory. Considerable progress has been made in recent months on the development of a theory of propagation at short distances over very irregular terrain such as is often encountered in cities. However, although this theory is able to explain fairly well the results obtained by the Columbia Broadcasting System in their color television trials in the 500–700 Mc band, it is not expected that it can be applied to FM broadcasting since one of the essential assumptions in this theory would no longer apply at these lower frequencies. Probably the only satisfactory solution to the irregular terrain FM propagation problem will be achieved through a statistical study of field intensities recorded in mobile receiving stations. Extensive data of this kind for the present FM band are not available at this time. However, based on such measurements made in the vicinity of 40 Mc, it would appear that the ground wave fields, except in very hilly or mountainous terrain, may be expected to exceed one-fifth of the smooth earth ground wave values in about 99% of the receiving locations.[9] Consequently, if we increase the required field of 10 microvolts/meter by a factor of 5 and then determine the distance to the 50 microvolt/meter ground-wave contour using the smooth earth theory, we may expect about 99% of the listeners at that contour to obtain a satisfactory FM service, provided an outside half-wave antenna and a sensitive receiver are used for reception, and provided the terrain is not unusually hilly or mountainous.

When the terrain is hilly or mountainous, an improved estimate of the expected fields can often be made by using a different radius of the earth which more nearly simulates the average of the terrain between the transmitting and receiving antennas. This method will be explained later.

VI. The Systematic Effects of Terrain and of Tropospheric Ducts

At distances beyond 20 to 30 miles, in addition to the effects of irregular terrain, irregularities in the lapse rate of the refractive index of the lower atmosphere cause the received fields to vary from minute to minute, hour to hour, and seasonally, the amount of the field intensity variations increasing with increasing distance and decreasing somewhat with an increase in the antenna heights. In an effort to learn more in detail about the magnitude of these variations, continuous measurements have been made at the National Bureau of Standards of the field intensity of FM broadcast station WCOD at Richmond, Va., a distance of about 100 miles. During the period of these recordings from June 10 to August 8, 1947, this station operated on 96.3 Mc with an effective power of 4.6 kw in the direction of Washington. Fig. 5 shows the profile of the transmission path between Richmond and Washington. The details on an expanded height and distance scale of the profile at each end of the path are given in the upper part of this figure. The lower profile corresponds to the entire path. These profile diagrams have been drawn with an effective earth's radius equal to four-thirds of its actual value in order to make allowance for the systematic effects of air refraction. The profile of the transmission path is of great importance in connection with estimating field intensities in the FM broadcast band. At great distances the received field intensities are dependent upon the curvature of the transmission path as determined by the ground profile, relative to the curvature of the radio ray path which, in turn, is determined by the lapse rate with height of the refractive index of the atmosphere. The curvature of the transmission path may be determined in the manner shown on this figure. Remembering that the radio waves are propagated through the

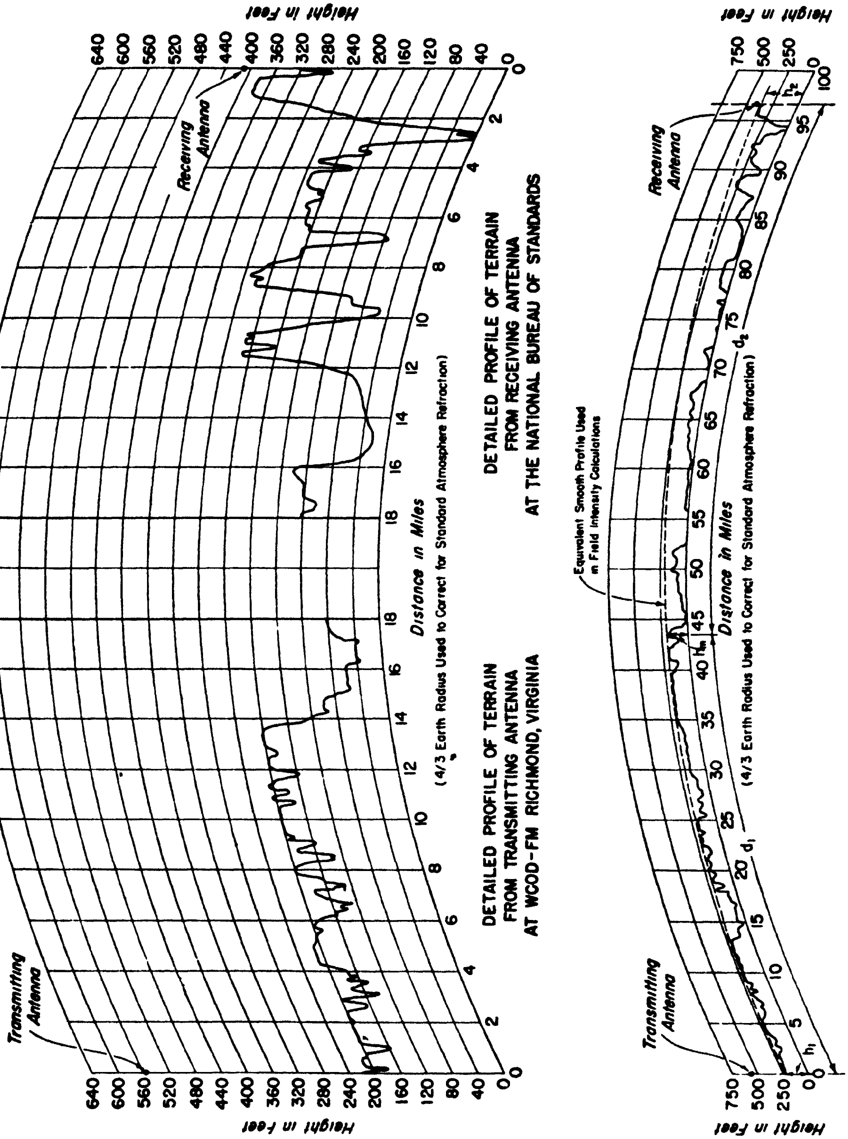


FIG. 5.—Terrain profile WCOD-FM Richmond, Va. to the National Bureau of Standards.

earth very poorly, it seems plausible that the effective value for the curvature of the earth to be used in estimating the curvature of the transmission path may be found by determining the arc of a circle (shown as a dashed line on the figure) which just grazes the earth at three points along the path, namely, at the two ends and at some point between. The curvature of this arc would thus be the effective curvature of the earth relative to which the radio transmission path is to be calculated; this can be determined by simple geometry from the heights above sea level, h_1, h_2, and h_m of the three points along the transmission path, together with the distances, d_1, d_2, from the ends of the path to the intermediate point.

Thus, if we write $k_p a$ for the radius of curvature of the equivalent smooth profile relative to which the actual transmission path is measured, in which a is the radius of the earth and thus the radius of curvature of the reference smooth spherical surface relative to which h_1, h_m, and h_2 are measured, then it is easy to prove that the ratio, k_p, may be expressed:

$$\frac{1}{k_p} = 1 - \frac{2a}{d}\left(\frac{h_1 - h_m}{d_1} + \frac{h_2 - h_m}{d_2}\right) \tag{1}$$

If we express the three distances d, d_1, and d_2 in miles and the three heights h_1, h_2, and h_m in feet, then eq. (1) becomes:

$$\frac{1}{k_p} = 1 - \frac{1.5}{d}\left(\frac{h_1 - h_m}{d_1} + \frac{h_2 - h_m}{d_2}\right) \tag{2}$$

In the particular case of this Richmond-Washington profile, both h_1 and h_2 are larger than h_m and the effective curvature of the earth for this transmission path was found to be about 6% less than the average curvature of the earth; i.e., $1/k_p = 0.94$. It will be noted that the terrain near the Washington end of the path lies well below this effective spherical surface. This would not be expected to influence to any great extent the received field intensities over this path, provided the bending due to air refraction is sufficiently small so that the curvature of the radio ray path is less than the curvature of the transmission path. It should be noted that the transmitting and receiving antenna heights are to be measured above this equivalent smooth profile. If the profile were lowered at the receiving end of the path the transmission path curvature would be increased but the expected fields would decrease only a little since this increased curvature would be largely offset by the fact that the effective receiving antenna height would then be larger. Presumably the proper position for the equivalent smooth profile is the one corresponding to the highest calculated values of the field intensity. This proper position would, in general, be expected to depend upon the amount of bending of the radio rays due to air refraction. Such variations have not been included in the theoretical fields to be shown for this transmission path, the transmitting and receiving antenna heights being taken, for purposes of calculation using the smooth earth theory, as fixed at 360 feet and 30 feet, respectively, above the equivalent smooth profile shown on Fig. 5, and thus corresponding to the fixed value of $k_p = 1.06$.

Air refraction almost always has the effect of bending the radio rays downward so that the relative curvature between the ray path and the transmission path is reduced. Under some circumstances, this downward curvature of the radio ray path may be as great as or actually greater than the curvature of the equivalent smooth profile. In this case, the received fields would be expected to be as great as or greater than the fields over a flat earth. Fig. 6 shows the field intensities of FM broadcast station WCOD as recorded on three successive days in August. On August 4, the station came on the air about 6:25 in the morning, the fields gradually increased throughout the day until a little after midnight, at which time the received field increased to a very much higher level and the fading, which had occurred at a fairly rapid rate during the day, decreased both in amplitude and frequency of occurrence. The arrow on this chart indicates the level of field intensity corresponding to propagation over a flat earth and we see that this level was exceeded for the half hour just prior to 1:00 A.M., at which time the station went off the air. Presumably this favorable propagation condition lasted throughout the night since the fields were again very strong the following morning when the station began broadcasting at 6:25 A.M. A plausible theoretical explanation for these strong fields accompanied by a comparative absence of fading is obtained if we assume that air refraction increased the curvature of the radio ray path at these times by an amount equal to or greater than the curvature of the transmission path. It is well known that the systematic effects of air refraction may be included in ground wave propagation calculations by using an effective radius of the earth k times its actual value a. Thus if we consider that the radio frequency refractive index of the air, n, decreases at a uniform rate (dn/dh) with the height, h, above the earth, then the appropriate value of k may be determined by the following equation:

$$\frac{1}{k} = 1 + \frac{a}{n} \cdot \frac{dn}{dh} \tag{3}$$

For a standard atmosphere $(dn/dh) = -12 \times 10^{-6}$/thousand feet so that $k = \frac{4}{3}$; but the value of (dn/dh) required to increase the curvature of the radio ray path until it equals or exceeds the curvature of the earth

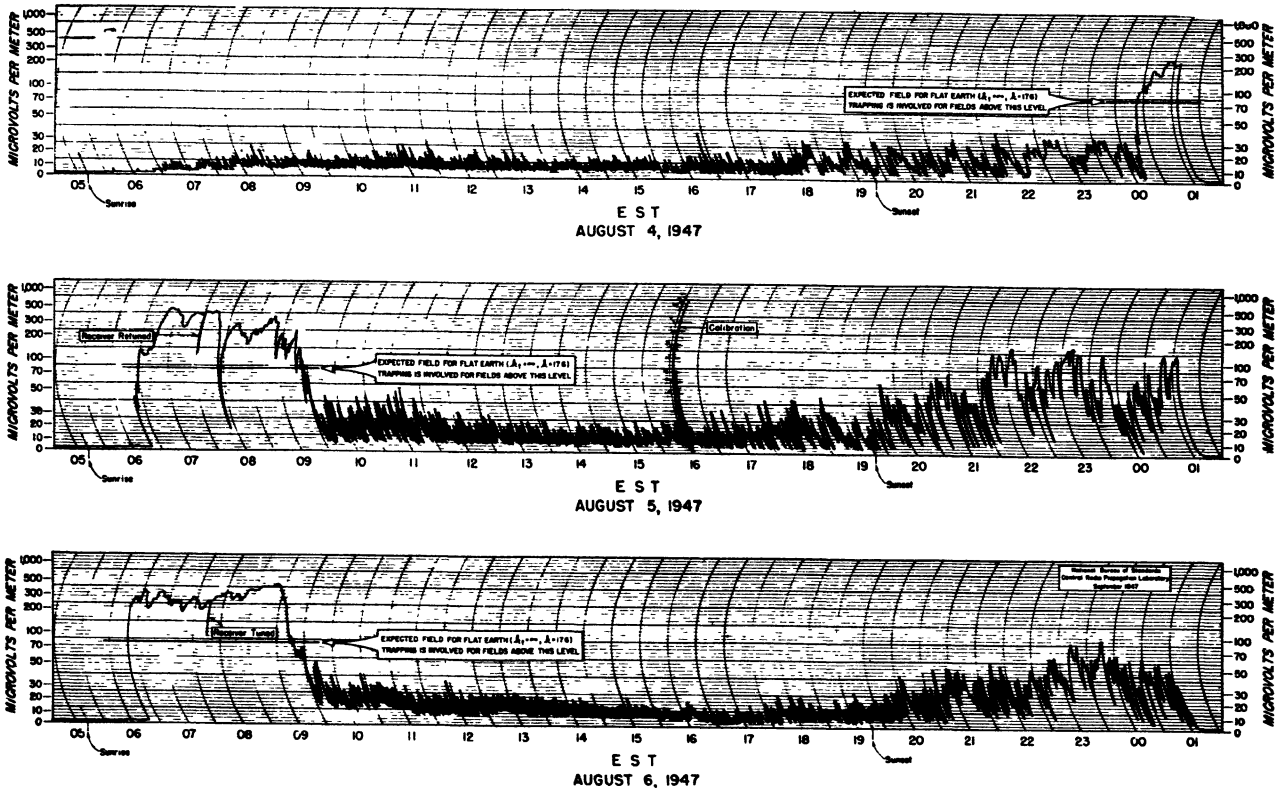


FIG. 6.—Field intensity of FM broadcast station at Richmond, Va. recorded at Washington, D. C. Frequency 96.3 Mc; antenna power 2.3 kw; antenna power gain = 2 relative to a half wave dipole; distance 96.6 miles; transmitting antenna height 360 feet above local terrain; receiving antenna height 30 feet above local terrain.

is four times as large, i.e., $k = \infty$ when $(dn/dh) = -48 \times 10^{-6}/$ thousand feet. Meteorological studies indicate that the radio frequency refractive index of the air cannot be expected to have this large a lapse rate throughout an extended height interval. However, lapse rates even larger than this frequently do occur over small height intervals and, at the high radio frequency here involved, these large changes in refractive index have an effect on the propagation which is equivalent to a smaller change in refractive index occurring over a larger height interval. When these large changes in refractive index occur near the surface of the ground, a surface duct is formed which guides the radio waves around the curved earth. The theory of the guiding action of such ducts was developed during the war by several research workers in this country and in England, and the results presented on Fig. 7 are based upon the work of H. G. Booker, who was in the Telecommunications Research Establishment in England during the war and is now at Cambridge University.[10] Fig. 7 shows the effectiveness of atmospheric ducts of various widths in reducing the curvature of the radio ray path or, what amounts to the same thing, in increasing the effective radius of the transmission path. The inset on this figure shows the meteorological characteristics of the surface duct for which Booker obtained a solution of the ground wave propagation problem; thus, this inset shows the refractive modulus, M, as a function of height. We see by Fig. 7 that, for atmospheric ducts with this shape, the required duct widths for trapping (k infinite) are greater than 1170 feet at 100 Mc; larger values of duct widths are required at lower frequencies for the same values of k in proportion to the factor $(100/f_{mc})^{\frac{2}{3}}$. The use of the ordinary ground wave theory[8] using the values of k determined by this figure, for calculations of the fields to be expected in the presence of a duct does, of course, provide only approximate results since, although it provides an accurate determination of the attenuation with distance, the height-gain functions will be increasingly inaccurate for the larger values of k. It is possible to obtain better results by using the more nearly exact solutions of this problem which are available in the literature.[10,11,12] However, the engineers in the Technical Information Division of the Federal Communications Commission have discovered that the use of large values of k in the ordinary ground wave theory gives results which are in good agreement with the high values of field intensity observed at VHF frequencies for small percentages of the time. Consequently, it seems likely that the connection shown here between large values of k in the ordinary ground wave theory and the theory of propagation in ducts will be of considerable use in understanding FM propagation. It is desirable to emphasize, of course, that resort to these more elaborate duct theories is

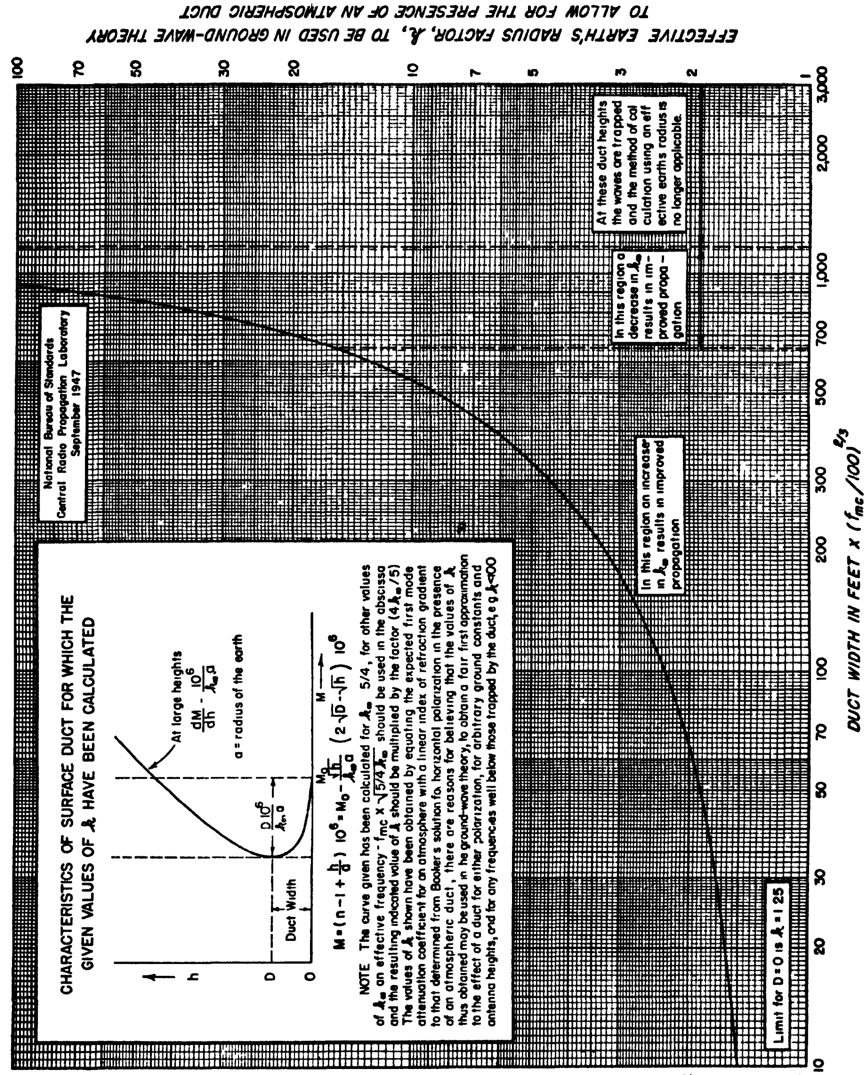
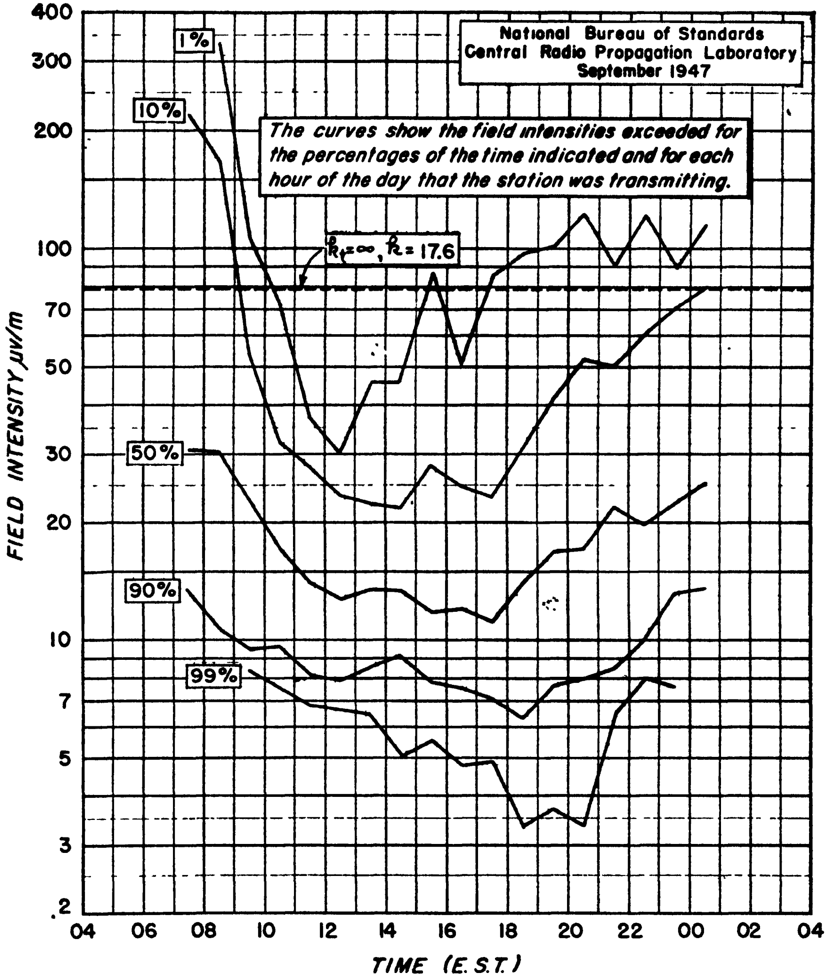


FIG. 7.—Expected field intensities in the presence of a Booker atmospheric duct may be calculated approximately by using spherical earth ground wave theory and an effective radius of the earth k times the actual value. (The effect of k is given explicitly in the paper by K. A. Norton, *Proc. Inst. Radio Engrs*, **29**, 623–639, Dec. 1941 as well as in F. E. Terman's Radio Engineering Handbook, McGraw Hill, New York, 1943.)

necessary only when the distances are large, and then only for that portion of the time when effective atmospheric ducts are present along the transmission path. For the overland propagation paths which are


National Bureau of Standards
Central Radio Propagation Laboratory
September 1947

The curves show the field intensities exceeded for the percentages of the time indicated and for each hour of the day that the station was transmitting.

FIG. 8.—Distribution of field intensities received at Washington, D. C. from FM broadcast station WCOD at Richmond, Va. from June 10 to August 8, 1947, inclusive. Frequency 96.3 Mc; antenna input power 2.3 kw; antenna power gain = 2 relative to a half wave dipole; distance 96.6 miles; transmitting antenna height 360 feet above local terrain; receiving antenna height 30 feet above local terrain.

usually involved in frequency modulation broadcasting effective atmospheric ducts would be expected after the sun sets and the earth begins to cool the atmosphere. Under favorable circumstances, this cooling may continue throughout the night with the result that a duct of great width is formed and the received fields would then be expected to reach their peak values early in the morning before the sun has had an opportunity to warm the earth and destroy the duct. The fields received over the Richmond-Washington path have been observed to have that general behavior. This is illustrated in Fig. 8 which shows the field intensities exceeded for various percentages of the total time during the period of recording from June 10 to August 8 for each hour of the day that the station was transmitting. The fields exceeded the flat earth level (as indicated by the dashed line at 79 microvolts/meter) for more than 10% of the time in the morning up to 9:00 A.M. and for more than 1% of the time up until 10:30 A.M. In the evening the fields exceeded the flat earth value for more than 1% of the time after 5:20 P.M. and exceeded this level for as much as 10% of the time during the latest hour for which recordings were available.

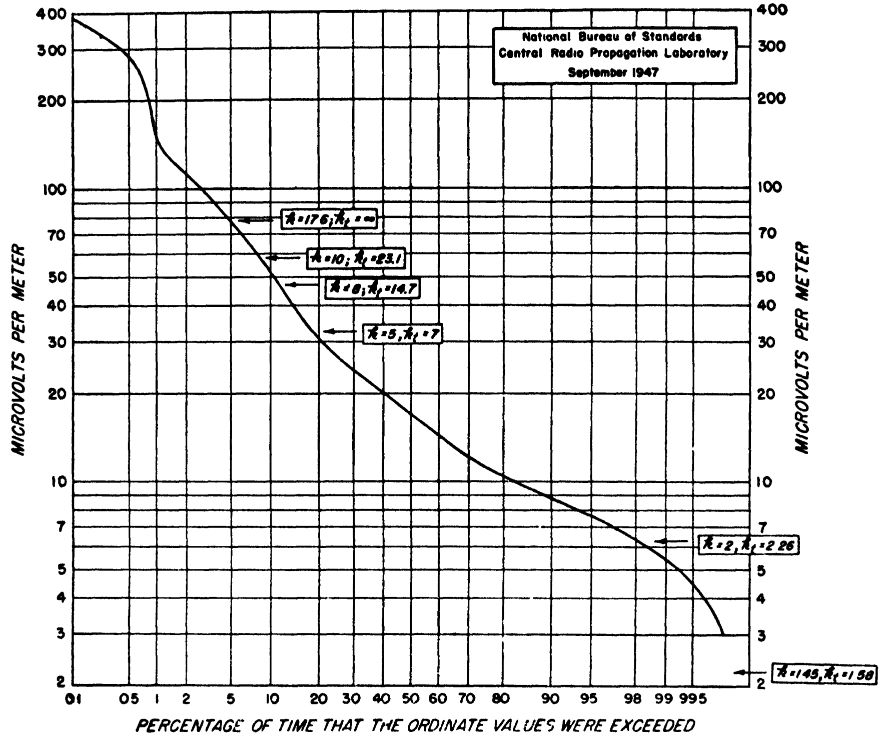


FIG. 9.—Distribution of field intensities received at Washington, D. C. from FM broadcast station WCOD at Richmond, Va. from June 10 to August 8, 1947, inclusive. Frequency 96.3 Mc; antenna input power 2.3 kw; antenna power gain = 2 relative to a half wave dipole; distance 96.6 miles; transmitting antenna height 360 feet above local terrain; receiving antenna height 30 feet above local terrain.

Fig. 9 shows the distribution of field intensities received over the Richmond-Washington path for all of the hours for which recordings were made. It should be noted that the 79 microvolts/meter to be expected over the path for a flat earth was exceeded for about 5% of this total time. The 99% field was 5.4 microvolts/meter, indicating that an increase in antenna input power up to only 8 kw is all that would be necessary to bring this up to the 10 microvolts/meter considered to be required for 99% of the time for a satisfactory FM broadcast service in quiet rural areas. The signals received from Richmond were usually observed to be of broadcast quality, especially in the early morning hours, even in the presence of the rather high man-made noise level at the Bureau.

The effective radius of the earth to be used in field intensity calculations, when the curvature of the transmission path is different from that of the curvature of the earth, may be obtained by subtracting the curvature of the radio ray path from the curvature of the equivalent smooth profile and then using the resulting radius of curvature as the effective radius of the earth in the calculations. If we write k_t for the ratio of this resultant radius of curvature to that of the earth, then this effective value, k_t, may be determined by means of the following equation:

$$\frac{1}{k_t} = \frac{1}{k_p} + \frac{1}{k} - 1 = 1 + \frac{a}{n} \cdot \frac{dn}{dh} - \frac{2a}{d}\left(\frac{h_1 - h_m}{d_1} + \frac{h_2 - h_m}{d_2}\right) \tag{4}$$

This effective value, k_t, is the appropriate value to be used in the ground wave theory[8] to include in the calculations the systematic effects of both air refraction and irregular terrain. On Fig. 9 k_t denotes this ratio for the Richmond-Washington path and k denotes the corresponding value this ratio would have had, for the atmospheric condition considered, if the transmission had been over a path with the same curvature as that of the earth. Thus, for a standard atmosphere $k = \frac{4}{3}$, the corresponding value for the Richmond-Washington path is $k_t = 1.44$ and the expected field over this path is 1.6 microvolts/meter. This value is not shown on the figure since the received field exceeded this value for nearly 100% of the time. The value of k required for flat earth propagation over this path is only 17.6 compared to the infinite value which would have been required over a transmission path for which the equivalent smooth profile has a curvature equal to that of the earth. Thus, we see that comparatively small changes in the propagation path profile have the effect of modifying considerably the atmospheric gradients required for a given field intensity. A value of $k = 17.6$ corresponds, by Fig. 7, to a duct width of $675 \times (100/96.3)^{\frac{2}{3}} = 692$ feet on 96.3 Mc and we conclude, since the flat earth fields over this path were exceeded for 5% of the time that atmospheric ducts with effective widths in excess of 690 feet were present for as much as 5% of the time. Adequate meteorological data are not available to substantiate this conclusion but such duct widths are not unreasonable; however, for the particular changes in refractive index implied by this particular model of refractive index vs. height (as defined in the inset on Fig. 7) a duct width of 690 feet corresponds to a decrease of 26 M units from the surface up to the height at which the minimum occurs and this is probably much too large. This would suggest that some other model for the refractive index distribution might be more appropriate;[10,11,12] however, the presently available information on the meteorology of the troposphere is insufficient for a quantitative solution to this problem in any case and this is the reason for our adoption of the simple parameter k as a measure of the systematic bending due to air refraction, instead of resorting to the more elaborate although more exact duct theories.

VII. THE TROPOSPHERIC WAVES RESULTING FROM REFLECTION AT ATMOSPHERIC BOUNDARY LAYERS

From what has been shown so far, it might be assumed that the received fields in the FM band at great distances can always be explained in terms of a difference in the curvature of the radio ray path and the transmission path. Unfortunately this is not the case, as we see on Fig. 10 where we have shown field intensity vs. distance curves for 46.7 Mc and for various assumed values of k_t varying from 1.2 up to an infinite value, the latter corresponding to flat earth propagation. Also shown on this chart are the field intensities of Station WABC-FM in New York City, exceeded for various percentages of the time, as measured at Princeton, N. J.; Andalusia, Pa. and Laurel, Md. The data at Princeton were obtained by RCA while the data at the other two receiving locations were obtained by the Federal Communications Commission. Two things should be noted. First, the high 1% measured fields are in the neighborhood of the field expected over a flat earth. Second, the 90 and 99% fields as measured at Andalusia lie far below the theoretical ground wave curve for $k_t = 1.2$ which is the lowest value consistent with the known characteristics of the lower atmosphere. It might be assumed that this behavior of the 90 and 99% fields could be explained as being due to terrain irregularities along this particular path, but, aside from the fact that there is nothing unusual about this path, measurements made at similar distances on other paths demonstrate that the 99% field always lies below the expected ground wave fields in this intermediate range of distances. Fig. 11 presents a set of theoretical curves which do have the proper shape to agree with the experimental data. Five theoretical

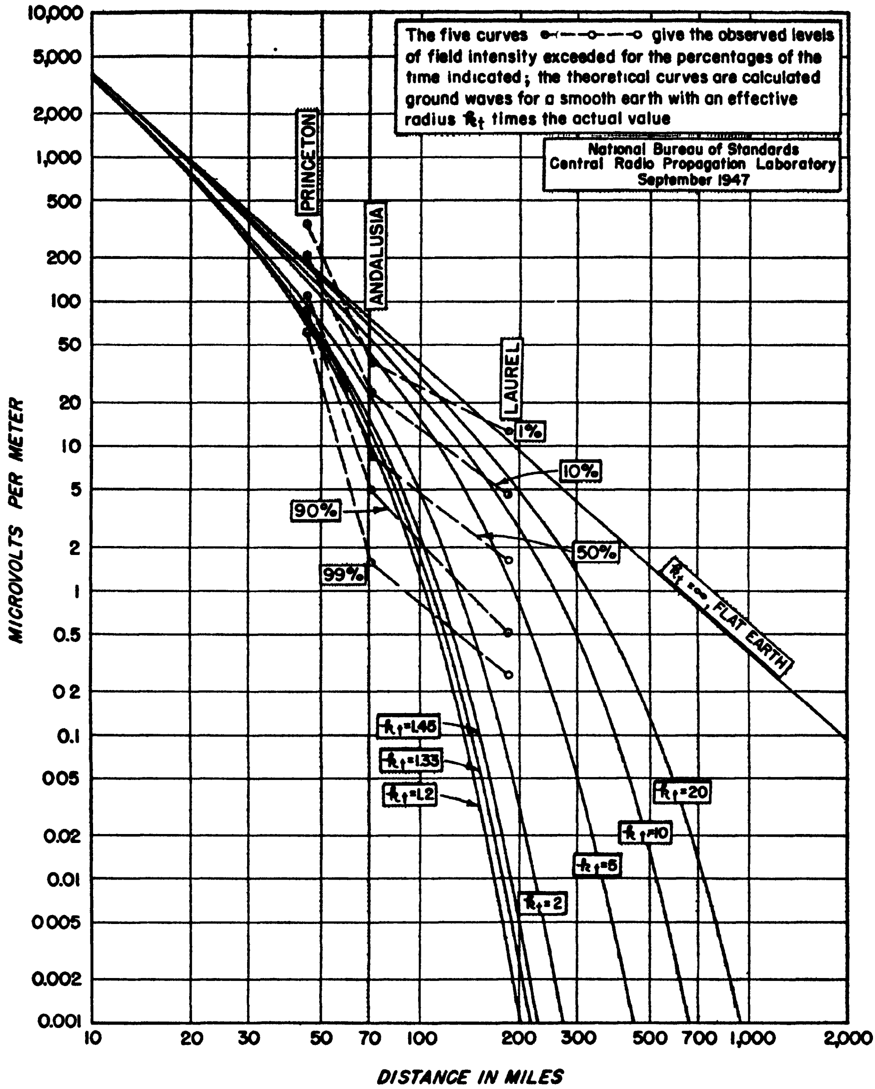


FIG. 10.—Field intensity of FM broadcast station WABC-FM in New York City recorded at Princeton, N. J.; Andalusia, Pa. and Laurel, Md. from Aug. 8 to Sept. 13, 1945; 271 simultaneous hours of recording between 6–10 A.M. and 3–11 P.M. inclusive. Frequency 46.7 Mc; data and theoretical curves based on 1 kw radiated from a half wave dipole; transmitting antenna height 780 feet above local terrain; receiving antenna height 30 feet above local terrain.

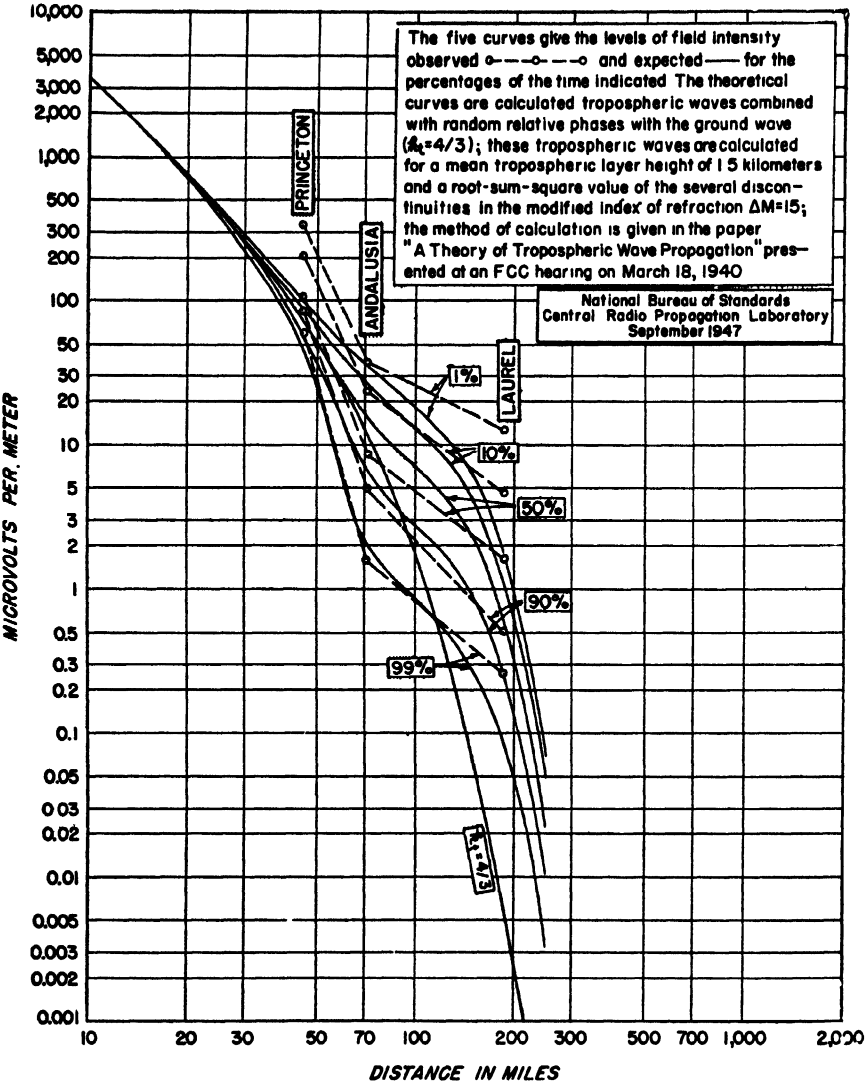


FIG. 11.—Field intensity of FM broadcast station WABC-FM in New York City recorded at Princeton, N. J.; Andalusia, Pa. and Laurel, Md. from Aug. 8 to Sept. 13, 1945; 271 simultaneous hours of recording between 6–10 A.M. and 3–11 P.M. inclusive. Frequency 46.7 Mc; data and theoretical curves based on 1 kw radiated from a half wave dipole; transmitting antenna height 780 feet above local terrain; receiving antenna height 30 feet above local terrain.

curves are given corresponding to the five percentages of the time 1%, 10%, 50%, 90% and 99% for which the experimental data are given. These theoretical curves were obtained by means of a tropospheric wave theory, the details of which were presented at a Federal Communications Commission hearing on March 18, 1940.[13] According to this theory, these tropospheric waves are reflected at various levels in the troposphere at which there are assumed to be more or less abrupt discontinuities in the distribution with height of the index of refraction.

This tropospheric wave theory[13] is very similar to the theory commonly used for calculating the intensities of downcoming ionospheric waves;[14] thus the spherical earth ground wave theory is used for calculating the intensities of the waves to be expected at the point of reflection in the troposphere. Next a coefficient of reflection at the troposphere is determined and finally, by assuming a new source at the point of reflection in the troposphere, the spherical earth ground wave theory is again used for calculating the attenuation of the reflected waves during propagation from the troposphere to the receiving antenna.

In calculating the coefficient of reflection at the troposphere the discontinuities in index of refraction are assumed to occur over a height interval very much smaller than the wavelength in which case the reflection coefficient is independent of the frequency. Epstein[15] has solved this reflection problem for changes occurring over a finite height interval and this solution is shown graphically in a recent paper by Smyth and Trolese.[16] As might be expected, the intensity of reflection decreases at the shorter wavelengths as soon as the wavelength becomes comparable to or smaller than the height interval involved in the discontinuity of the index of refraction; however, when conditions are favorable for large amounts of systematic bending, i.e., large values of k, the angles of incidence at the various tropospheric layers may become less than the critical angle[15] and the reflection will become complete irrespective of the frequency. There is some experimental evidence that these tropospheric wave-reflection coefficients are smaller at the higher frequencies; for example, the field intensities received from FM broadcast stations at very large distances (greater than 150 miles) are stronger in the 40–50 Mc band than in the 88–108 Mc band.[17] This difference is most readily explained by assuming that the discontinuities, ΔM_s, in index of refraction occur over an interval of height of the order of 10 feet or more and that the individual values of ΔM_s are too small to produce total reflection at the angles of incidence ordinarily involved:[18] under these conditions the tropospheric layer reflection coefficient would decrease in intensity with increasing frequency in this frequency range.

In the F.C.C. paper referred to[13] methods of calculation were given for the waves reflected from single discontinuities at particular heights. A study of meteorological data indicates that many small discontinuities in refractive index are simultaneously present in the atmosphere, being distributed over a wide range of heights from the earth's surface up to heights of ten or twenty thousand feet. Reflected tropospheric waves may be expected from each of these discontinuities at various distances along the propagation path and, since the lengths of the individual propagation paths for these various waves will differ by several wavelengths we may expect these individual tropospheric wave components to arrive at the receiving point with random relative phases. The heights of the discontinuities are probably continually varying and this will cause the relative phases of the separate wave components to change with time; thus the resultant downcoming tropospheric wave (the vector sum of the component waves) will vary in amplitude over a wide range. This is believed to be the cause of the rapid changes in intensity which occur from minute to minute in a downcoming tropospheric wave. The distribution of the intensity of such a tropospheric wave with time is given by the Rayleigh distribution.[18,19,20] If we write P for the percentage of time that the field intensity, E, of the resultant downcoming wave is exceeded, we obtain by the use of Lord Rayleigh's theory:

$$P = 100e^{-(E/E_t)^2} \tag{5}$$

$$Ee^{-i(\omega t+\alpha)} = E_0\Delta M_1 e^{-i(\omega t+\alpha_1)} + \cdots + E_0\Delta M_i e^{-i(\omega t+\alpha_i)} + \cdots + E_0\Delta M_m e^{-i(\omega t+\alpha_m)} \tag{6}$$

$$E_t^2 \equiv (E_0\Delta M)^2 = (E_0\Delta M_1)^2 + \cdots + (E_0\Delta M_i)^2 + \cdots + (E_0\Delta M_m)^2 \tag{7}$$

In the above α_i denotes the phase of the i^{th} component wave and is assumed to be random relative to the others; E_o denotes the expected tropospheric wave field intensity for a single discontinuity of refractive index $\Delta n = 10^{-6}$, so that $E_o\Delta M_i$ is the intensity of the i^{th} component which is assumed to have been reflected at a discontinuity of amount $\Delta n = \Delta M_i \cdot 10^{-6}$. We see by eq. (7) that, according to the Rayleigh theory, ΔM is the root-sum-square value of the m discontinuities ΔM_i. It may be shown[19] by means of eq. (5) that E_t is the root-mean-square value of the resultant downcoming wave and that the median value of this downcoming wave, $E_{50\%} = 0.8326\ E_t$. Eq. (5) provides an accurate representation of the intensity distribution of a composite wave, E, as defined by eq. (6), whenever m is large ($>$ 4 or 5) and when each component wave is much smaller than the root-sum-square value of all of the waves, i.e., $(\Delta M_i)^2 \ll (\Delta M)^2$. Such conditions would be expected in tropospheric wave propagation problems in most cases. However, on the West Coast of the United States a very large and more or less permanent reflecting tropospheric layer has been discovered[16] and the value of ΔM_i corresponding to this layer might frequently be larger than the root-sum-square value for the coexisting smaller layers; calculations in this case can be handled in the manner described below for adding a ground wave to a Rayleigh distributed tropospheric wave.

At great distances, such that the mean square value of the downcoming tropospheric wave, $E_0 \Delta M$, is large compared to the ground wave, E_g, the above theory may easily be generalized to give the expected over-all distribution of the intensity of a ground wave plus the tropospheric waves. Thus we simply add a term $E_g e^{-i(\omega t + \alpha_g)}$ to eq. (6) and a term E_g^2 to both sides of eq. (7) so that eq. (5) becomes:

$$P = 100 e^{-E^2/(E_t^2 + E_g^2)} \quad (E_g \ll E_t) \tag{8}$$

Since the simple Rayleigh distribution (8) is applicable only at large distances, and since no expression, comparable in simplicity to eq. (8), is available when E_g is comparable to or larger than E_t, the solution to this latter problem is given graphically in Fig. 12.

Using the methods referred to[8,13] for computing E_g and E_t, and the above theory for combining these waves with random relative phase, the five theoretical curves shown in Fig. 11 were obtained; these correspond to a root-sum-square value of the several discontinuities in the index of refraction $\Delta n = 15.10^{-6}$, these discontinuities being assumed to occur over a range of heights with a mean value of 1.5 km. This value of Δn is required for a height interval small compared to the wavelength; larger values would be required if the changes occur over a larger height interval.[15,16] The curves for the various percentages of the time were obtained by combining with random relative phases the ground wave for a value of $k = \frac{4}{3}$ with the individual waves from the several discontinuities. The good agreement between the experimental data and the theoretical curves for the 90 and 99% levels is quite evident. Still better agreement could have been obtained by calculating the expected tropospheric waves for several different layer heights, and then adding these resulting waves with random relative phase. Since the reflecting tropospheric layers are known to occur over a range of heights, the higher ones being more favorable for long distance propagation, and conversely, the lower layers favoring shorter distance propagation, the use of a combination of layer heights would increase the tropospheric wave intensities somewhat, relative to the values given both at small and at large distances, thus providing an improved fit to the experimental data. It will be noted that the 1% measured field lies well above the 1% theoretical tropospheric layer field. This is to be expected since we have already seen in Fig. 10 that these high fields are very probably due to a large increase in the systematic bending of the ground wave. Still further evidence in favor of this assumption is the fact that the fading, which would be expected with tropospheric waves, is not present to any marked extent when the received fields are very strong. This may be

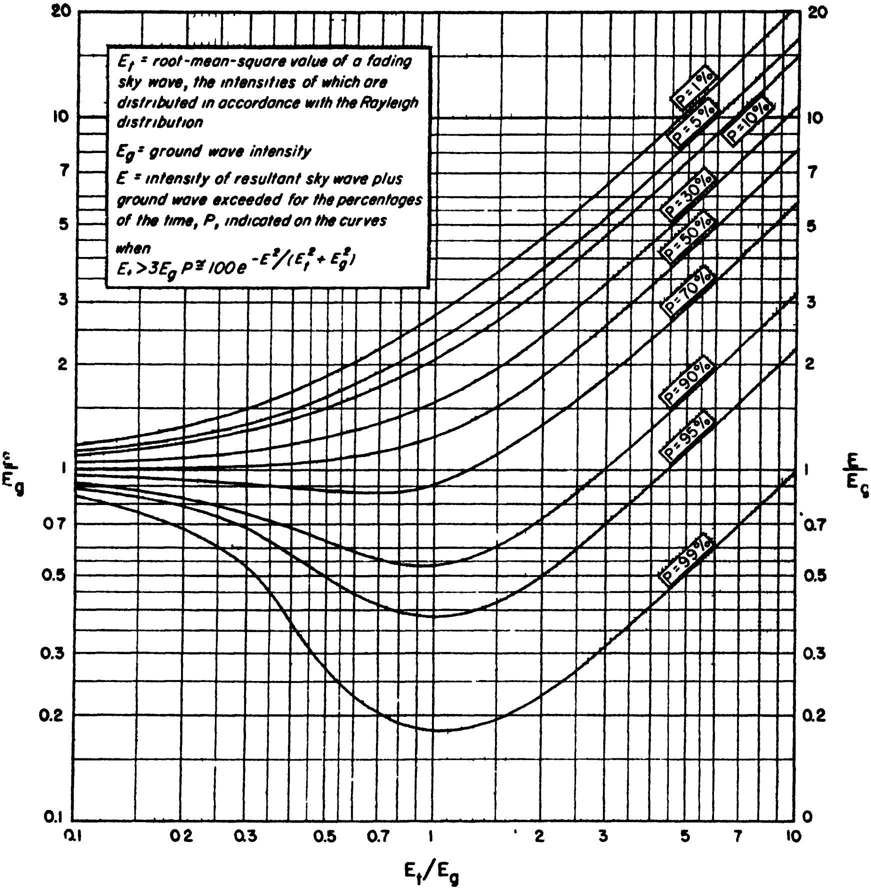


FIG. 12.—The expected distribution of the intensity of the resultant of a ground wave combined with a Rayleigh distributed sky wave.

understood by noting on Fig. 12 that the fading range is small when E_g becomes large relative to E_t.

VIII. THE COMBINED EFFECTS OF DUCTS AND OF RANDOM TROPOSPHERIC WAVES

Fig. 13 shows the fields to be expected in summer months on 98 Mc as calculated for a smooth earth using a combination of the theories outlined above. The results on this figure are for a transmitting antenna height of 1000 feet and a receiving antenna height of 30 feet. It is the 99% field which determines the service range of an FM station and we see that this field decreases very rapidly with distance between 40 and 70 miles. Similar results are given on Fig. 14, but now for a transmitting antenna height of only 100 feet. In this case, the fading occurs at much shorter ranges, the precipitous drop in field intensity occurring now in the range from 20 to 30 miles. There are several ways in which the theoretical curves given on Figs. 13 and 14 may be improved. For

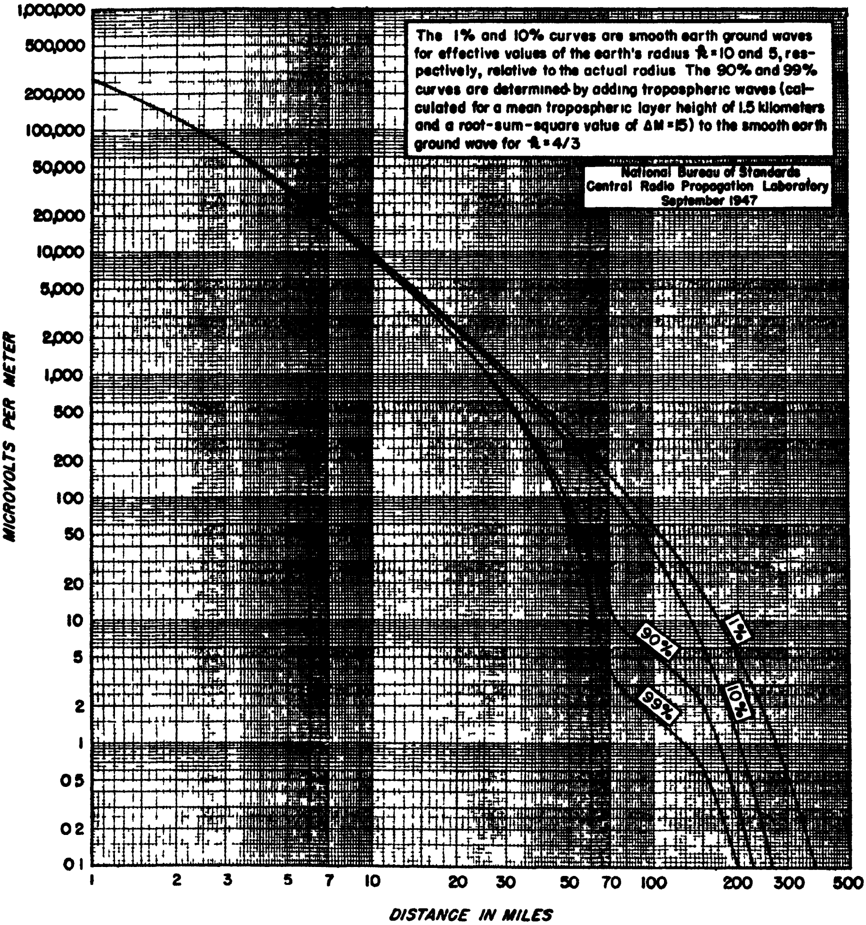
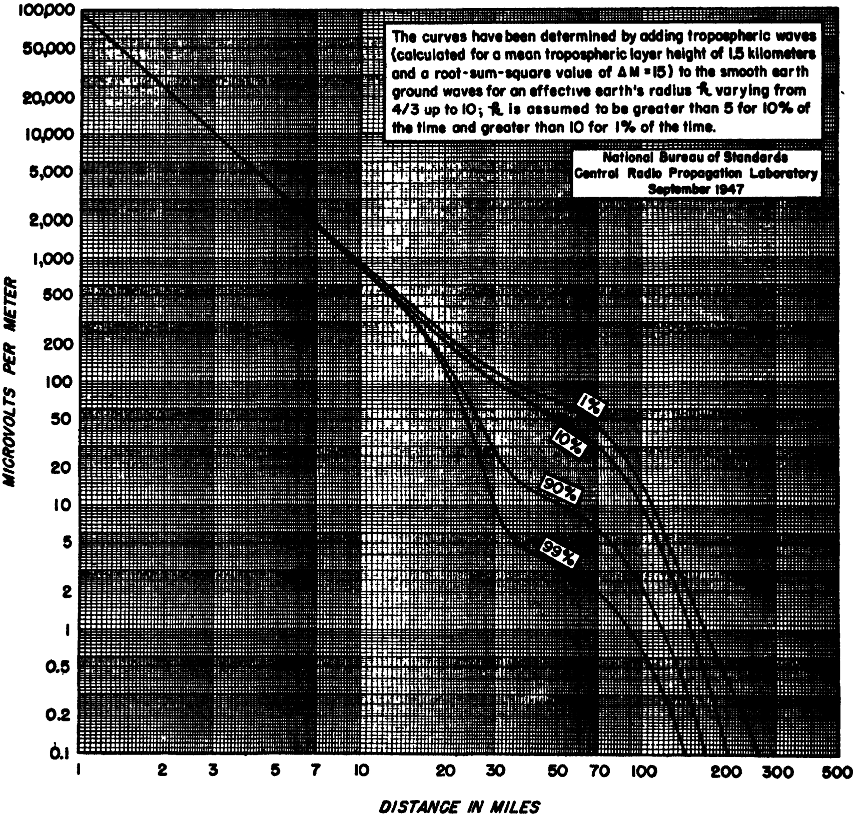


FIG. 13.—Field intensity *vs* distance expected in summer months on 98 Mc for the percentages of the time shown Transmitting antenna height 1000 feet, receiving antenna height 30 feet; 1 kw radiated from a half wave dipole.

example several tropospheric layer heights should be used simultaneously as already suggested; also the intensities of the tropospheric waves should be calculated for various values of k so as to include the effects of systematic bending on the tropospheric wave intensities. Until such time


The curves have been determined by adding tropospheric waves (calculated for a mean tropospheric layer height of 1.5 kilometers and a root-sum-square value of $\Delta M = 15$) to the smooth earth ground waves for an effective earth's radius k varying from 4/3 up to 10; k is assumed to be greater than 5 for 10% of the time and greater than 10 for 1% of the time.

National Bureau of Standards
Central Radio Propagation Laboratory
September 1947

FIG. 14.—Field intensity *vs.* distance expected in summer months on 98 Mc for the percentages of the time shown. Transmitting antenna height 100 feet; receiving antenna height 30 feet; 1 kw radiated from a half wave dipole.

as more experimental data become available such refinements in the theory are probably unwarranted.

IX. THE CALCULATED SERVICE AND INTERFERENCE RANGES OF FM BROADCAST STATIONS

Fig. 15 gives the expected summer month service ranges for FM broadcasting stations as determined by the theoretical curves on Figs. 13 and 14. The summer month conditions are given since the variations in

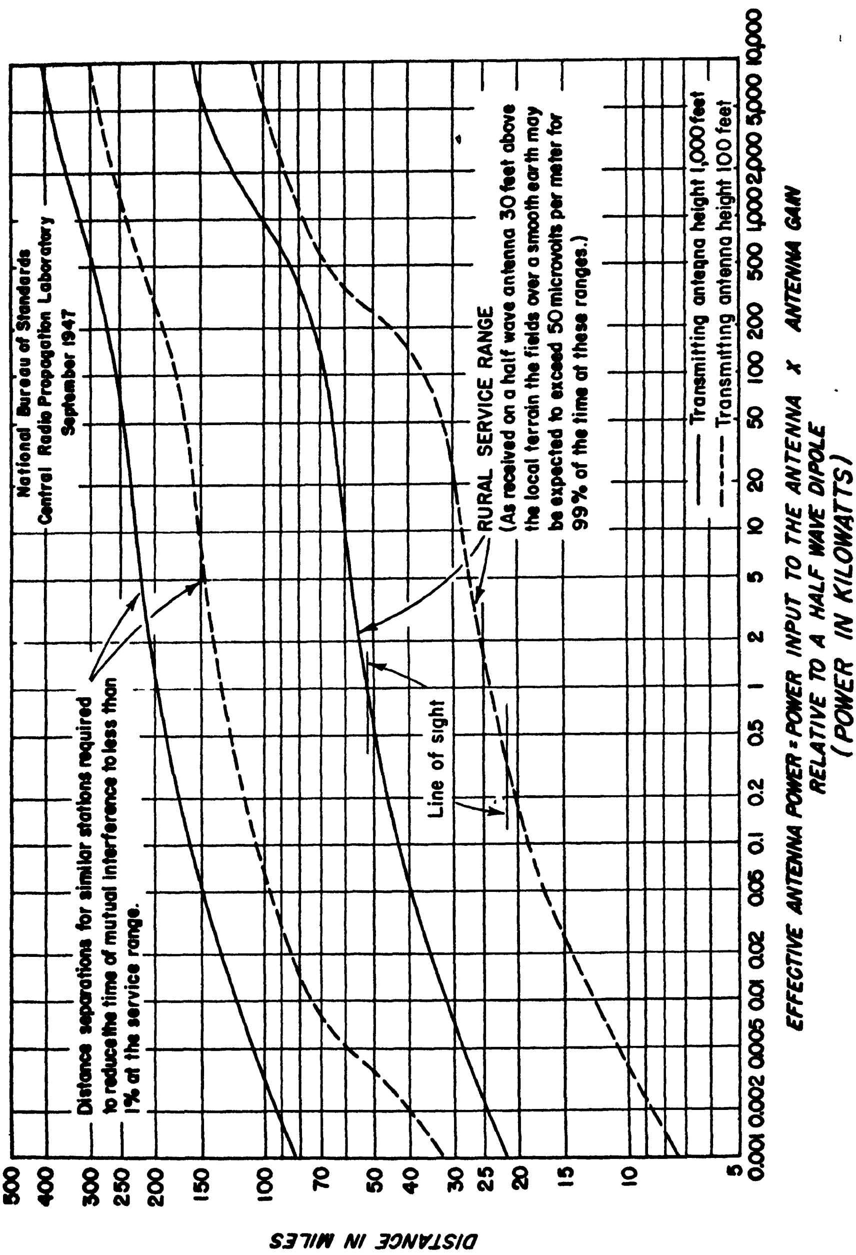


FIG. 15.—Summer month service ranges for FM broadcast stations. Also shown are the separations, between stations operating on the same channels and with the same effective power and antenna height, required to prevent interference at the service range for more than 1% of the time.

the received fields are greatest during this period. During the winter months, the air is systematically much drier and this has the effect of greatly reducing the radio propagation variations occurring during these months. We see by this figure that service to a distance corresponding to line of sight is possible with an effective power of about 1 kw. Effective powers up to as much as 1000 kw can probably be obtained with facilities now available by using an antenna with a power gain of 20 in conjunction with a 50 kw transmitter; the use of this much power will, as shown on Fig. 15, provide a satisfactory service to distances several times the distance to the horizon. The service ranges shown are the expected 50 microvolt/meter contours of the field intensity exceeded for 99% of the time at a height of 30 feet above a smooth earth. The use of the 50 microvolt/meter smooth earth field rather than the 10 microvolts/meter actually required in rural locations makes allowance for a 5 to 1 reduction in field intensity from the smooth earth values which may be expected in some locations due to irregularities in terrain along the transmission paths.[9] The service ranges are given for two transmitting antenna heights, the dashed curve being for a 100-foot height while the solid curve corresponds to a 1000-foot transmitting antenna height.

Also shown on this chart are the separations between identical co-channel stations required to reduce the time of mutual interference between these stations at their 50 microvolt/meter service contours to less than 1% as determined from the theoretical curves on Figs. 13 and 14. In determining these required distance separations, interference was considered to exist whenever the desired smooth earth field was less than 10 times the undesired smooth earth field; the use of a factor of 10 rather than the factor 2, which is sufficient on some receivers for rejecting an undesired cochannel FM signal, makes allowance for a 5 to 1 deviation from the calculated smooth earth ratio of the fields which may occur in some locations owing to differences in the irregularities in terrain along the transmission paths from the desired and undesired stations.[9] In determining these required distance separations it was also assumed that there would be no correlation in the fading between the desired and the undesired signals. Some correlation undoubtedly does exist with either high or low fields occurring together rather than independently for the desired and undesired signals so that the distance separations shown may be slightly greater than would actually be required.

Those familiar with the FM allocation practices of the Federal Communications Commission will notice that the distance separations here indicated as being necessary have not always been maintained in practice. This is not necessarily undesirable and merely means that a higher level of field intensity is being protected against interference. This may be understood best by noting that a similar increase in power for both the desired and the undesired stations will not change the ratio of their fields at any receiving location and thus will not modify the areas within which each station is able to provide service free from interference for 99% of the time; an increase in power for both stations will simply increase the field intensity at this service contour in proportion to the square root of the ratio by which the power is increased. Consequently, if a determination is desired for the distance at which a 1000 microvolt/meter field intensity may be expected for 99% of the time for a 20 kw station with a 1000-foot transmitting antenna height, the distance is determined on Fig. 15 corresponding to an effective power lower in proportion to the square of the field intensity ratio $(\frac{1000}{50})^2 = 400$, i.e., for 20 kw/400 = 0.05 kw; the expected 1000 microvolt/meter contour for such a 20 kw station is, by Fig. 15, about 40 miles, and a similar 20 kw station at a distance of 150 miles would not be expected to cause interference at the 1000 microvolt/meter contour for more than 1% of the time. Fig. 15 may also be used to determine the field intensity contour and distance at which interference-free service may be expected for 99% of the time when the distance to an interfering station operating with the same effective power and antenna height is known. Thus, suppose that such an interfering station is operating with an effective power of 20 kw with a 1000-foot antenna height at a distance of 200 miles; this distance corresponds, by Fig. 15, to a power of 1.25 kw so that a field intensity of $50 \times \sqrt{20/1.25} = 200$ microvolts/meter is being protected against interference, and interference-free service may be expected for 99% of the time out to a distance of 53 miles.

X. THE EFFICIENT ALLOCATION OF FACILITIES TO FM BROADCAST STATIONS

If we define the most efficient utilization of a single FM channel to be that corresponding to an allocation of stations which will be capable of serving the maximum percentage of a given area in which it is desired to furnish radio service, then it can be shown that all of the stations allocated to that channel should be placed at equal distances from each other on a triangular lattice, the required optimum separation being shown on Fig. 15. Furthermore, each station should have the same power and antenna height, and these should be as large as it is possible to make them. The reason for the latter statement is not obvious but can be determined from the data on Fig. 15. Thus, the total area of the triangular lattice is proportional to the square of the distance separation given on Fig. 15 while the total area served is proportional to the square of the service range also given on Fig. 15. The percentage of the total area served is therefore directly proportional to the square of the ratio of the service to separation distances given on Fig. 15. It will be noted that this ratio increases with increasing antenna height and with increasing power. As an example of the much more efficient channel utilization possible with high powered stations employing high antennas rather than lower powered stations with lower antennas a determination was made of the percentages of a given large area which it is possible to serve in several cases. These percentages are given in the table. It is important to notice that the efficiency of the allocation, as measured by the percentage of area capable of being served by a single channel, increases with increasing antenna height much more rapidly than it increases with power. The efficient utilization of FM channels would thus appear to be promoted best by the utilization of the highest transmitting antennas available. In this connection, it would appear that "stratovision," i.e., the system of broadcasting involving transmission from aircraft cruising in the stratosphere, might well offer considerable advantages.

TABLE I. Idealized allocation of similar FM stations for serving a large area such as the United States.

Effective power (kw)	1		10		100		1000		10,000	
Antenna height (ft.)	100	1000	100	1000	100	1000	100	1000	100	1000
Service range (mi.)	24	52	28	60	36	67	78	102	108	157
Separation distance (mi.)	130	198	151	228	176	256	240	320	303	410
Percentage of total area capable of being served by a single channel (%)	12	25	13	25	16	25	38	37	46	78
Approximate number of channels required to provide a single service to the entire United States	10	5	10	5	8	5	4	4	3	2
Approximate number of stations required to provide a single service to the entire United States	4110	889	3070	637	1810	535	486	273	228	83

XI. THE OPTIMUM FREQUENCY FOR AN FM BROADCAST SERVICE

In view of the controversy over the recent decision of the Federal Communications Commission to change the FM band from its former position in the spectrum, 42–50 Mc to its present position, 88–108 Mc, it is considered desirable to outline briefly some of the technical factors involved. In considering this problem it is necessary to take into account the effective service ranges to be expected when more than one FM station is permitted to operate on the same channel simultaneously. In fact the demand for FM facilities has been so great that it has sometimes been necessary to allocate stations to the same channel even at shorter separations than those indicated as desirable on Fig. 15. The justification for this practice has already been discussed; it appears to be a practical solution in those populous areas of the United States where the demand for facilities exceeds the number available. However, in less populous areas, where the economic situation is such that only a very few stations may be operated at a profit, it is desirable to choose a band of frequencies for FM such that the maximum possible *interference-free service areas* may be realized so as to provide the rural listeners, now most in need of an improved broadcast service, with the benefits of this new system of broadcasting. From this point of view the choice made by the Federal Communications Commission of the higher band of frequencies becomes quite clear.

The expected service and interference ranges for FM stations are shown as a function of the radio frequency (Fig. 16) for stations with an antenna height of 500 feet. It has been assumed on Fig. 16 that rural service will be available out to the distance over a smooth transmission path at which the field received on a 30-foot receiving antenna exceeds 50 microvolts/meter for 99% of the time. A field of 50 microvolts/meter has been assumed, rather than the 10 microvolts/meter actually required for service with typical receivers, in order to allow for the additional reduction in received field to be expected on some transmission paths as a result of irregularities in the terrain. The horizontal dashed lines on Fig. 16 represent the calculated service ranges for stations operating with effective powers of 1, 10, 100 and 1000 kw. The service ranges at 98 Mc were estimated from Fig. 15 by interpolation between the ranges given for 100 and 1000 foot antennas. These service ranges were then assumed to be approximately applicable throughout this frequency range since the theoretical and experimental data available at the present time are insufficient to determine just how much, if any, variation in range exists.

The available experimental data will be considered first. Carnahan, Aram, and Classen reported[21] measurements on 45.5 and 91 Mc made simultaneously over a 76-mile path between Richfield, Wis. and Deerfield, Ill. The transmitting antennas were at heights of 508 feet on 45.5 Mc and 468 feet on 91 Mc above the average elevation along the transmission path. The measurements were made continuously from 11 A.M. to 11 P.M. during the period from July 20 to September 21, 1945. Their

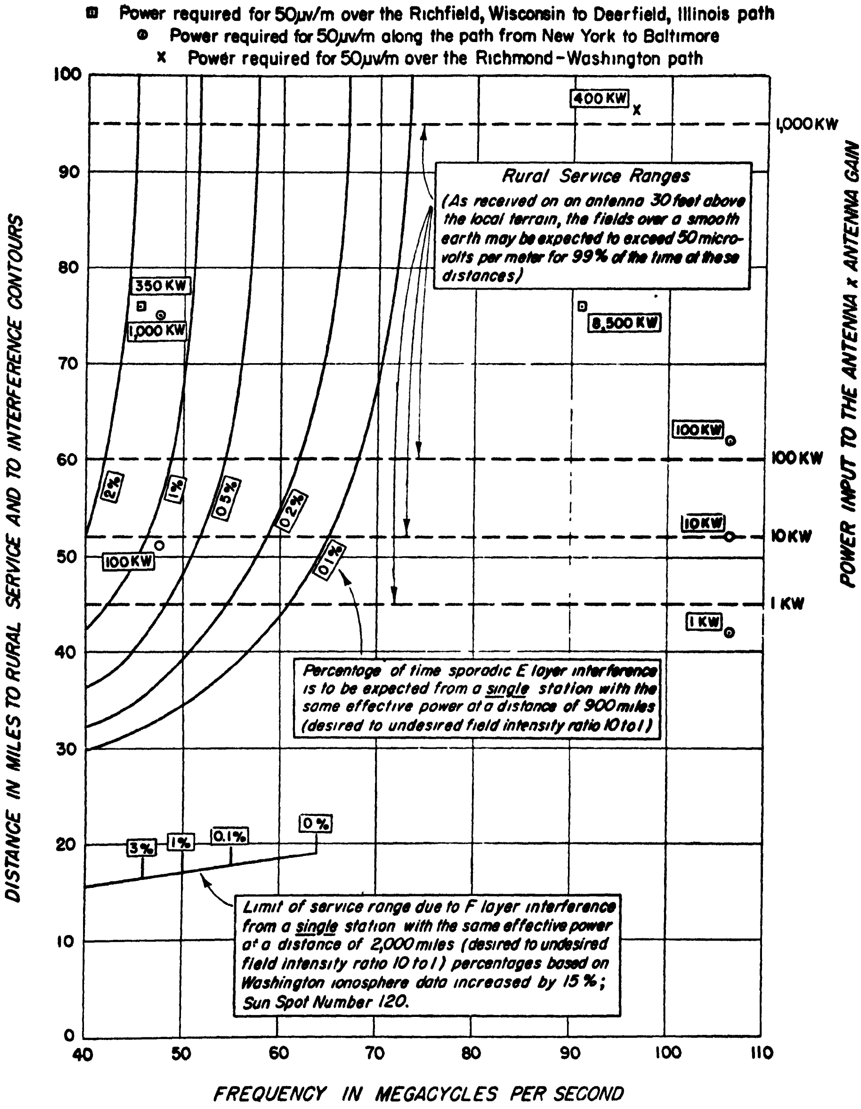


FIG. 16.—The effect of frequency on the expected rural service and interference ranges for FM broadcasting stations. Transmitting antenna height, 500 feet; receiving antenna height, 30 feet.

Fig. 10[21] shows that an effective power of 350 kw would be required at 45.5 Mc, and 8500 kw would be required at 91 Mc in order to provide a field of 50 microvolts/meter for 99% of the aforementioned period of time over this path. These two points are given on Fig. 16 for comparison with the calculated ranges.

E. W. Allen[17] has reported measurements made at Princeton, N. J.; Southampton, Pa. and Laurel, Md. of simultaneous transmissions on 47.1 and 106.5 Mc made available for this purpose by the Bamberger Broadcasting Service, Inc.; the distances to these three receiving stations were 45, 68 and 186 miles, respectively. The transmitting antennas for both frequencies were 570 feet above the street in New York City and the recordings were made daily except Saturdays from 8 A.M. to 12 midnight during the period June through October 1946. From the data shown on Fig. 5 of the Allen report[17] it is possible to estimate the distance at which 50 microvolts/meter would be exceeded for 99% of the time when either 100 or 1000 kw is used on 47.1 Mc and with 1, 10, or 100 kw on 106.5 Mc; these ranges are given on Fig. 16 for comparison with the calculated ranges.

Finally, from the data shown on Fig. 9 a determination can be made of the effective power required over the Richmond-Washington path to provide a 50 microvolt/meter field for 99% of the time at the Bureau of Standards receiving station; this was found to be 400 kw and is shown on Fig. 16. All of the points plotted on Fig. 16 represent measurements of field intensities over comparatively smooth transmission paths and consequently are suitable for direct comparison with the calculated service ranges indicated by the horizontal dashed lines. It is important to recognize that only one twenty-fifth as much power would be required over these particular transmission paths in order to provide the 10 microvolts/meter for 99% of the time, which is the actual field intensity necessary for service with typical receivers; thus, in particular, an effective power of only 16 kw would be sufficient to provide satisfactory service at 96.3 Mc over the 96.6 mile Richmond-Washington path. This extra factor of 25:1 in power has been introduced as already explained in order to provide service to those listeners who live in locations at the same distance but with more unfavorable conditions of terrain along the transmission path.

A study of Fig. 16 indicates the need for additional experimental data before safe conclusions can be drawn as to the reliable service ranges of FM stations; in particular, although the Richfield-Deerfield path measurements indicate that more power is required on the higher band for service at the same range, the New York to Baltimore data indicate that larger reliable service ranges are to be expected on the higher band when the same effective power is used. It is believed that at least a part of the observed differences on these two transmission paths is due to systematic differences in tropospheric propagation; thus, the Northeastern path is known to be characterized by slightly higher values of surface refractive index than the Midwestern transmission path, and may be subject to considerably different M gradients.

Since the available experimental data are inconclusive as to the expected variation of service range with frequency, it is desirable to see what light may be thrown on this question by the theory. Over a flat earth the received field intensity at great distances is directly proportional to the frequency; however, when obstacles, including the bulge of the earth itself, are present along the transmission path, additional attenuation is encountered and this is higher on the higher frequencies. Thus at distances far beyond the line of sight, the ground wave field intensities expected over a smooth spherical earth are actually greater on the lower frequencies.[8] However, two effects of the troposphere have been identified and these are expected, on theoretical grounds, to change in opposite directions as the frequency increases. Thus, surface ducts in the lower atmosphere are, according to Fig. 7, more effective at the higher frequencies in guiding the ground waves around the bulge of the earth. On the other hand the waves reflected at tropospheric boundary layers tend to be more completely reflected at the lower frequencies and this may result in stronger fields being received on the lower frequencies at the very great distances where these tropospheric waves are known to predominate; in this connection, however, it should be remembered that the propagation to and from the tropospheric boundary is often more favorable on the higher frequencies and this may result in a cancellation of the effects of the stronger reflection coefficients at the lower frequencies. The final conclusion to be drawn from these influences must await further data to evaluate the relative magnitude of these conflicting factors. Nevertheless, on the basis of the data and theories now available, it is the author's opinion that less power will be required on the higher frequencies on most transmission paths to produce the same field intensity (exceeded 99% of the time) at least up to distances of the order of 50 to 70 miles and, as will become evident in what follows, larger service ranges than this cannot be established on the lower band in any case because of long distance ionospheric interference.

It is rather surprising at first glance to note the much lower rate of attenuation of field intensity (exceeded 99% of the time) in the range from 60 to 95 miles as compared to the rate of attenuation in the range from 45 to 60 miles. Thus on Fig. 16 a 100 to 1 increase in power from 1 to 100 kw may be expected to increase the service range by only 15 miles whereas a subsequent increase of only 10 to 1 in power to 1000 kw increases the expected range by 35 miles. This more than 4 to 1 change in the rate of attenuation is due to a physical change in the mechanisms of propagation effective in these two distance ranges; thus, at the shorter ranges the received fields are principally ground waves, whereas at the larger ranges tropospheric boundary layer waves are involved. This reduced rate of attenuation is also evident on Fig. 15 where it may be seen to extend out to at least 150 miles. In analyzing 3- and 9-cm. propagation over the sea, Katzin, Bauchman and Binnian[28] have reported a similar change in the rate of attenuation at a range of 80 miles on 9 cm.; such a change was not observed, however, on 3 cm. It seems entirely possible that these higher frequency observations may be explained as being due to the same phenomena here described for FM propagation over land; the absence of the effect on 3 cm. would result from (1) the smaller tropospheric layer reflection coefficients to be expected for the shorter wavelength[15,16] and (2) the fact that the 3-cm. attenuation is already so much reduced by trapping that the small boundary layer reflections would be masked.

Before leaving the discussion of the service ranges, it should be noted that service has been defined in terms of the available field intensities expressed in microvolts/meter rather than in terms of microvolts available across the receiver input terminals. It has been argued by Dale Pollack in a discussion of a paper by E. W. Allen[22] that the receiver terminal voltage is a more appropriate measure since he considered that internal receiver noise rather than external noise is the factor limiting the reception and that half-wave dipole receiving antennas will provide only half as much voltage across the receiver terminals in the higher band when the same field intensity is available in both cases. However, the cosmic radio noise measurements reported by Herbstreit (pp. 347–380) show that external noise is the limiting factor in reception on frequencies less than 130 Mc with low noise figure receivers and, when these measurements are translated into field intensities required with a half-wave dipole receiving antenna, it may be seen on Fig. 4 that actually lower field intensities are required on the higher band for the same grade of service.

Also shown on Fig. 16 are the ranges at which interference would be expected on the lower frequencies from one other FM station operating on the same frequency at a distance of 900 miles and another station at a distance of 2000 miles. The interference from the latter would be propagated via the F layer of the ionosphere and, at frequencies less than the maximum usable frequency, would be expected to have an intensity approximately equal to the free-space field at this distance[28] since ionospheric absorption is negligible at these high frequencies. Thus the service range in the presence of F layer interference has been determined on Fig. 16 by determining the distance from the desired station at which the smooth earth ground wave field is 10 times the free space field from the undesired station. It should be noted that, so long as the same effective power is used by both the desired and undesired FM stations, the maximum distance at which interference-free service may be maintained is independent of the power used, since a similar increase or decrease in power for both stations would not change the ratio of the desired-to-undesired fields. The percentages indicated on this F layer interference contour are the percentages of the time throughout a typical sunspot cycle (with a smoothed maximum sunspot number of 120) that the maximum usable frequency would be expected to be sufficiently high to permit F layer interference over paths from points south of the border to points within the United States. These percentages were determined by increasing by 15% (to allow for the effect of latitude) the Washington, D. C. maximum usable frequencies exceeded for the indicated percentages of the listening hours (6 A.M. to midnight) as measured during the last sunspot cycle (1933–1944). Actual measurements of this F layer interference made on low band FM stations during 1946 and 1947 and reported by E. W. Allen[24] have confirmed that interfering fields even in excess of the free-space values are, as predicted,[23] often received and that this intense interference lasts for approximately the length of time anticipated from the maximum usable frequencies estimated from vertical incidence ionospheric soundings.

The interference to be expected from a low band FM station at a distance of 900 miles is attributed to sporadic E layer transmission and the field intensities to be expected in this case are much weaker. The interference ranges shown for 44.3 Mc were based on actual measurements made by the Federal Communications Commission of WGTR, Paxton, Mass., field intensities as received at Atlanta, Ga., a distance of 900 miles.[25,17] The values of sporadic E layer interference indicated for the other frequencies were extrapolated from these 44.3 Mc data by means of an empirical relation giving the percentage of time of occurrence of sporadic E reflections as a function of frequency as observed[26,27] at C.R.P.L.* using vertical incidence ionosphere soundings. The almost complete absence in the 88–108 Mc band of interference attributable to sporadic E layer transmission provides the justification, at least qualitatively, for the use of this method of prediction. Interference transmitted via the sporadic E layer is observed in the range of distances from about 400 to 1400 miles with maximum periods of occurrence at a distance of about 1000 miles and maximum field intensities at a distance of about

---

* **Central Radio Propogation Laboratory of the National Bureau of Standards.**

800 miles; these variations with distance and frequency are the principle identifying characteristics of this type of interference.

It has been argued by some that this F layer and sporadic E layer ionospheric interference occurs for such a small percentage of the time as to be of negligible importance. However, the interference shown on Fig. 16 corresponds to that expected from a single station operating on the same channel at the distances indicated. Actually, as has been pointed out in detail by E. W. Allen,[22] it is necessary as a practical matter to operate many stations simultaneously on the same and adjacent channels in order to accomodate the demand for FM facilities, and thus the percentages indicated on Fig. 16 would have to be multiplied in many instances by a factor of 5 or 10 for sporadic E interference and 2 or 3 for F layer interference. From the above analysis it becomes quite clear that much larger areas can be provided with an interference-free FM broadcast service by using frequencies in excess of 70 or 80 Mc.

As a final precaution it is desirable to emphasize that the service ranges shown on Figs. 15 and 16 are conservative in as much as they correspond to a 50 microvolt/meter field intensity rather than the 5 microvolts/meter actually required with an unusually good FM receiver as shown on Fig. 4. Consequently it may be expected that satisfactory reception for 99% of the time will be available to many rural listeners at distances 3 or 4 times the reliable ranges indicated on Figs. 15 and 16. It is desirable to point out that we are just beginning to learn a little about the characteristics of radio propagation in the FM band and much further experimental and theoretical research is indicated. One kind of data which is considered to be of utmost value in connection with such research is continuous recordings of the field intensities of FM and Television broadcast stations. Such data should be collected for a wide variety of meteorological and terrain characteristics in various parts of the country. It is hoped that the cooperation of individual broadcasters and of University research laboratories can be obtained in connection with such a measurement program. Until such time as these more complete data become available, it is hoped that the interpretations of FM propagation presented here may provide a rough guide for future research and for the day to day decisions which must be made by the broadcasters.

REFERENCES

1. Brunner, W. *Terr. Magn. Atmos. Elect.*, **44**, 247–256 (1939). These Zurich sunspot numbers are published for later years in subsequent issues of the above journal.
2. McNish, A. G., and Lincoln, J. V. Prediction of Annual Sunspot Numbers, Rept. No. CRPL-1-1, May 15, 1947, published by the Central Radio Propagation Laboratory, Natl. Bur. Standards, Washington, D. C.

3. "Basic Radio Propagation Predictions, CRPL series-D, a monthly publication giving ionospheric propagation conditions three months in advance as predicted by the Central Radio Propagation Laboratory, Natl. Bur. Standards; this monthly publication may be obtained by subscription from the Superintendent of Documents, Government Printing Office, Washington, D. C.
4. Waldmeier, M. *Terr. Magn. Atmos. Elect.*, **51**, 270 (1947).
5. Stewart, J., and Eggleston, F. *Astrophys. J.*, **91**, 72 (1940); *Phys. Rev.*, **55**, 1102 (1939); *Astrophys. J.*, **88**, 385 (1938).
6. Phillips, M. L. The Ionosphere as a Measure of Solar Activity, Rept. No. IRPL-R26, prepared in the Interservice Radio Propagation Laboratory, Natl. Bur. Standards, Washington, D. C.
7. Tilton, E. P. *Q.S.T.*, **31**, 58–61 (December 1947); **32**, 57–60 (January 1948). Norton, K. A. *Q.S.T.*, **31**, 13–17 (December 1947).
8. Norton, K. A. *Proc. Inst. Radio Engrs.*, **29**, 623–639 (1941); *Proc. Inst. Radio Engrs.*, **25**, 1192–1202 (1937).
9. Chapin, E. W., and Norton, K. A. Field Intensity Survey of Ultra High Frequency Broadcasting Stations, Report presented to the Federal Communications Commission Hearing in the Matter of Aural Broadcasting on Frequencies above 25,000 Kilocycles, March 18, 1940, F.C.C. Mimeo. No. 40004.
10. Booker, H. G., and Walkinshaw, W. The Mode Theory of Tropospheric Refraction and its Relation to Wave-Guides and Diffraction, pp. 80–127 of a report of a conference held on April 8, 1946: Meteorological Factors in Radio-Wave Propagation, published by the Physical Society and the Royal Meteorological Society, London, England.
11. Kerr, Donald E., Ed. Propagation of Short Radio Waves, Chap. 2, Radiation Laboratory Series, vol. 13, McGraw-Hill, New York, in press.
12. Pekeris, C. L. *J. Appl. Phys.*, **17**, 1108–1124 (1946).
13. Norton, K. A. A Theory of Tropospheric Wave Propagation, Report presented to the Federal Communications Commission Hearing in the Matter of Aural Broadcasting on Frequencies Above 25,000 Kilocycles, March 18, 1940, F.C.C. Mimeo. No. 40003.
14. Norton, K. A. A Theoretical Determination of the Intensity of Sky Waves at Intermediate Frequencies, Part III of a Supplementary Report on Wave Propagation for the C.C.I.R., April 1937, F.C.C. Mimeo No. 23743 or 84810.
15. Epstein, P. S. *Proc. Nat. Acad. Sci. U.S.A.*, **16**, 627–637 (1930).
16. Smyth, J. B., and Trolese, L. G. *Proc. Inst. Radio Engrs.*, **35**, 1198–1202 (1947).
17. Allen, E. W. Preliminary Report on East Coast Tropospheric and Sporadic E Field Intensity Measurements on 47.1 and 106.5 Mc. November 18, 1947. Report presented to the Federal Communications Commission Hearing on Sharing of Television Channels, Docket No. 8487.
18. Rayleigh, J. W. S. (Lord). *Phil. Mag.*, **10**, 73–78 (1880); see also Theory of Sound, 2nd ed., Macmillan and Co., London, paragraph 42a, 1894.
19. Norton, K. A. Discussion of Vernon D. Landon's *Proc. Inst. Radio Engrs.* paper, **30**, 426–429 (1942).
20. Norton, K. A. The Polarization of Downcoming Ionospheric Radio Waves, May 1942. Report prepared at the Federal Communication Commission in connection with a Natl. Bur. Standards project sponsored by the National Defense Research Committee. F.C.C. Mimeo. No. 60047.
21. Carnahan, C. W., Aram, N. W., and Classen, E. F. *Proc. Inst. Radio Engrs.*, **35**, 152–159 (1947).

22. Allen, E. W. *Proc. Inst. Radio Engrs.*, **35**, 128–152 (1947).
23. Norton, K. A. The Nature of Very-High-Frequency Ionospheric Wave Propagation, Report presented as Classified Exhibit No. 6 in the Hearing Before the Federal Communications Commission in the matter of the Allocation of Radio Frequencies Between 10 kc and 30,000 Mc, Docket 6651, F.C.C. Mimeo. No. 81009.
24. Allen, E. W. Measurements of F2 layer Propagation at 40–50 Mc, Report presented to the Federal Communications Commission on November 10, 1947 in the Hearing in the Matter of Amendments to the Commission's Rules and Regulations Governing Sharing of Television Channels, Docket No. 8487, F.C.C. Mimeo. No. 14026.
25. Allen, E. W. VHF Radio Field Strength Measurements: 1943–1944, Report dated September 28, 1944 and presented to the Federal Communications Commission in the Hearing on Docket 6651, F.C.C. Mimeo. No. 77785.
26. Phillips, M. L. Radio Propagation Conditions, Publication of the Interservice Radio Propagation Laboratory, Natl. Bur. Standards, Washington, D. C., October 14, 1943, p. 4 and February 14, 1944, pp. 3–4.
27. Phillips, M. L. *Trans. Am. Geophys. Union*, **28**, No. 1, 71–78 (1947).
28. Katzin, Martin; Bauchman, R. W., and Binnian, William. *Proc. Inst. Radio Engrs.*, **35**, 891–905 (1947).

Electronic Aids to Navigation

J. A. PIERCE

Cruft Laboratory, Harvard University, Cambridge, Mass.

CONTENTS

Page
I. Introduction 425
  1. Pilotage 425
  2. Dead Reckoning 426
  3. Fixing 426
  4. Navigation 426
II. Prewar Methods 427
  1. Direction Finding 427
    a. Loop Antenna 427
    b. Night and Shore Effect 428
    c. Automatic Direction Finding 429
  2. Azimuth Finding 429
    a. Radio Range 429
    b. Orfordness Beacon 430
    c. Cathode Ray Direction Finding 430
    d. Spaced Loop Direction Finding 430
III. Wartime Developments 431
  1. Elektra 431
  2. Sonne 431
  3. Radar 432
  4. Gee 432
  5. Loran 433
  6. Low Frequency Loran 433
  7. Decca 434
IV. Postwar Proposals 434
  1. Omnirange 435
  2. Teleran 435
V. Considerations of Range and Accuracy 436
  1. Classification of Systems 436
  2. Ground- and Sky-wave Range 436
  3. Range of Pulse Systems 439
  4. Maximum Accuracy 440
  5. Geometric Factors 443
  6. Range vs. Accuracy 448

I. INTRODUCTION

1. Pilotage

The fundamental duty of a navigator is the selection of a route to be followed by his vessel. His choice may depend upon the presence of hazards or may require only the knowledge of present and desired positions and of potential drifts to be expected from motion of the sea or air. The end point of the route is the destination, not often chosen by the navigator, but his attention may often be fixed upon the route to a way point or check point selected by himself.

It is axiomatic that the navigator must know his present position and his destination, at least, before he can choose a route, and from that his course and the heading of his vessel. In the simplest and most important form of navigation, pilotage, the vessel is directed in terms of landmarks seen by the navigator (in this case often called the pilot) and the knowledge of present position is not usually explicit. In navigation out of sight of landmarks the navigator customarily finds his position, usually in terms of latitude and longitude or distance and direction from a known point, and consults a chart to find the relation between his position and his destination.

2. Dead Reckoning

A position, once known, can be carried forward indefinitely by keeping continuous account of the direction and speed of the vessel. This process is called dead reckoning. Unfortunately, heading and speed are not often capable of precise measurement and the effects of drift can only be approximated, so that, on the average, the errors of dead reckoning increase in proportion to the length of time it has been carried on. Ordinarily a dead reckoning position is seriously in error after the vessel has traversed a few hundred miles without any external indication of position.

3. Fixing

In contrast to dead reckoning, a fix is a determination of position without reference to any former position. The simplest fix stems from the observation of a recognizable landmark and, in a sense, any fix may be thus described. A celestial fix is nothing but the recognition of the unique point from which, at a given instant, a number of stars appear at their observed altitudes. Similarly, most radio aids to navigation present a certain indication at only one point, or only a few points, on the surface of the earth. Landmarks beyond the range of vision can be recognized by radar or by the fathometer.

4. Navigation

Since a fix is fundamental and dead reckoning is derived, it is natural that most electronic aids to navigation are aids to the determination of fixes or, like radar, are devices that extend the range at which pilotage may be carried on. Thus the dividing line between pilotage and "long range" navigation is somewhat nebulous. The best distinction arises from the method of determining position: if explicit, it is conventional navigation; if implicit, it is pilotage.

Now navigation does not consist of the determination of position or establishment of a compass heading to be followed by a helmsman. Navigation requires the exercise of judgment; it is a choice (based on all available data concerning position, destination, weather, natural and artificial hazards, and many other factors) of one out of many courses of procedure that may lead to the required result. There can be, therefore, no electronic navigational *system*, but only aids to navigation. An aircraft or ship may be made to follow automatically a predetermined course by the use of equipment that performs the dead reckoning function, or may be made to follow a line of position known to pass through a desired objective. Neither of these achievements constitutes navigation by the equipment. The automatic devices simply extend the control exercised by the navigator in time or space.

II. Prewar Methods

1. Direction Finding

With these facts in mind it becomes worthwhile to examine the radio aids to navigation available in 1939. These were all directional devices, in the sense that they measured angles subtended at a radio transmitter or receiver, but we should distinguish carefully between direction finding (the measurement of the relative bearing of a fixed radio transmitter as "seen" from a receiver on a moving vehicle) and azimuth finding (the measurement of the true bearing or azimuth of the vehicle as seen from a fixed station). At short distances, of course, the two results differ only in the sense of the absolute direction but the techniques and uses of the methods differ greatly. There are, in fact, three main classes of directional aids that should be examined separately:

a. Direction finding. Transmitters fixed and receiver on the vehicle.
b. Azimuth finding. Receivers fixed and transmitter on the vehicle.
c. Azimuth finding. Transmitters fixed and receiver on the vehicle.

The fourth expected class, the first with the direction of transmission reversed, is actually indistinguishable from the second.

a. Loop Antenna. Direction finding, in its simplest form, is accomplished by rotating a loop antenna until a signal received by it is reduced to zero when the plane of the loop is normal to the direction from which the signal comes. If the received signal is linearly polarized, and if there are no inhomogeneities in the conductivity of the region near the loop, the indication of direction is very accurate. Unfortunately these conditions are not often attained in practice. The received signal is likely to consist, at least in part, of one or more sky-wave components that have, by the action of the ionosphere, had their polarization altered into elliptical form. The effect upon the direction finder is an error, shifting from time to time as the polarization changes, that may approach 180° and is frequently of the order of 10–15°. Masses of metal near the antenna distort the received field so that the resultant of the original signal and the reflections from local objects has an apparent direction that usually differs significantly from the true direction to the transmitting station. This effect is constant for a given true direction so that it can be corrected, but the constancy of the corrections depends upon constancy of the geometrical conditions; thus on shipboard, for instance, the corrections are not often accurately known.

Another defect of the simple loop is that the amplitude of the signal passed through the receiver varies sinusoidally with rotation. Thus a sharp indication is had only when the amplitude is at or near zero. When the loop is oriented for this signal condition noise can, of course, be received from other directions. The indication of direction is thus at its best when the signal-to-noise ratio is least favorable.

b. Night and Shore Effect. In addition to these inherent qualities of the loop antenna, simple direction finding suffers from the fact that it is not the direction to the transmitter that is observed but rather the orientation of the wave front received from the transmitter. This may be significantly different in two major cases. The most common is when the signal is predominantly received by sky-wave transmission and is reflected from a sloping layer. This condition is most probable in the case of north-south transmission near sunrise or sunset, and is then called "night effect." The other important case is that of ground-wave transmission near and more or less parallel to a shore line. The velocity of propagation is significantly smaller over land than over water and the signals are transmitted along paths that are somewhat concave toward the land rather than radial from the transmitter. Either this "shore effect" or the night effect may be responsible for errors of several degrees.

The first attempt to avoid part of the difficulties inherent in direction finding for shipboard use lay in the erection of groups of shore-based direction-finding stations. These stations, usually three in number, would take bearings on the signal from a shipboard transmitter upon request. The bearings would be plotted at one of the shore stations so that a fix could be reported to the vessel within a few minutes. This technique has the great advantage that the local distortions of the received fields can be kept small and constant and that complex and precise equipment can be used. The requirement of a number of special stations, however, has militated against extensive use of this system so that in recent years its most important use has been for detection of the location of enemy vessels.

c. Automatic Direction Finding. A modification of the loop receiving technique has been used, especially in aircraft automatic direction finders, to avoid the difficulty of detecting a weak signal in the neighborhood of the null. This concept consists in the use of two loop antennas at right angles with identical receiver channels leading to some form of indicator. As the two loops are rotated together, the signals in the two channels become equal only when the apparent direction of the station bisects the 90° angle between the planes of the two loops. As implied above, this or a similar technique is required for the automatic direction finders that have been used extensively in recent years. An error in the orientation of the crossed loops leads to a difference in the two outputs, a difference whose relative amplitude measures the magnitude of the error, while the channel having the stronger signal identifies the sense in which the error should be corrected. It is therefore easy to use a simple servomechanism that will maintain the equality of the signals in the two channels by rotating the loops to the correct relative bearing.

2. *Azimuth Finding*

a. Radio Range. Another series of attempts to improve direction finding led to the development of the first azimuth finders, the Radio Range and the Orfordness beacon. The Radio Range is very nearly the crossed loop direction finder in reverse. A pair of loops are used as transmitting antennas and are disposed nearly at right angles so that their radiation maxima lie in alternate quadrants. The loops are driven alternately from a common transmitter with an interlocked keying sequence so that one loop repeatedly radiates the morse letter A in the silent intervals between the emissions of the letter N from the other loop. Thus along the line where the two loops provide equal signal strength a steady tone is heard. Departures from the equisignal zone are marked by reception of A's or N's, depending upon the sense of the departure. Since each loop antenna has two radiation maxima, there are thus provided four equisignal zones, or ranges, that may be mutually perpendicular or may be established in any four directions by adjustment of the antenna characteristics.

The Radio Range thus provides four well-established lanes that may be made to coincide with desired routes. In regions outside the equisignal zones the only information received is the identification of the sector containing the navigator. The great merit of the system lies in the elimination of the local quadrantal errors and of all measuring or adjusting at the navigator's position. The signals may be received on any antenna and the data are gathered simply by listening.

b. Orfordness Beacon. The Orfordness beacon also enjoys the elimination of quadrantal errors but requires some measurement by the navigator. In return for this effort he is able to determine his azimuth from the transmitting station at any point rather than along four lines only. The transmitter uses a single loop antenna that is rotated exactly once per minute. Thus the sharp minima of the loop pattern sweep chrough the navigator's position every 30 seconds. The signal is shut off for a short interval just before one of the minima passes through North and is started when this "negative beam" is exactly North. With a stop watch the navigator times the interval from the start until one of the minima is heard. This interval, computed at 6°/second, measures the azimuth (or reciprocal azimuth) of the navigator as seen from the transmitter. It is thus a true azimuth measuring system, although the 180° ambiguity must be resolved by conventional navigational methods.

c. Cathode Ray Direction Finding. There are two important techniques in direction finding that, while not particularly new, were first used extensively and successfully during the recent war. The first, both chronologically and in wartime importance, was the use of crossed loops with identical amplifying channels and a cathode ray indicator. This truly instantaneous direction finder was of immense value in locating the sources of extremely short bursts of high speed radiotelegraph signals transmitted from German submarines. In other respects the technique has little advantage over the classical methods, except that the uncertainty of the directional observations, due primarily to polarization errors, is perhaps a little more obvious than with other indicators.

d. Spaced Loop Direction Finding. The other recent development of an earlier idea is the spaced loop direction finder. This is more properly the spaced antenna direction finder, but loop antennas are ordinarily the most convenient. In this arrangement two spaced antennas, connected through identical amplifying channels to a phase comparator, are rotated bodily about an axis midway between them until the phases of the two received signals become identical. If the physical separation of the antennas is only a small fraction of a wavelength, as it frequently must be, the sensitivity of the phase comparison must be high. The method, however, has the unique advantage that the observed direction does not depend upon the composition and polarization of the incoming signals. The precision of measurement is therefore limited primarily only by mechanical accuracy, by the phase sensitivity, and by the homogeneity of the space surrounding the equipment. Thus it closely approaches the ultimate accuracy, because direction finding can only measure the apparent direction of the resultant signal which is made up of the actual signal received from a distance (and this may not have travelled over a great circle path) and the signals reflected from neighboring regions where there are discontinuities in the conductivity or refractive index.

III. Wartime Developments

1. *Elektra*

The only significant German wartime contribution to radio aids to navigation was the development of Sonne and its parent system, Elektra. Elektra is essentially a multiple radio range, providing a large number (frequently 24) of equisignal zones. Deviations from the equisignal zones were detected by hearing dots on one side of the zone and (interlocked) dashes on the other. Other methods, such as direction finding on the transmitting station had to be employed to solve the ambiguity problem. The beauty of the system was the provision of such a large number of identifiable lanes with a simple three-tower antenna system of good efficiency so that, for the first time, highly accurate azimuth finding could be carried on at ranges of several hundred miles. The azimuths of several equisignal lanes could be adjusted (together, not independently) so that one of them would extend along any chosen great circle from the transmitter. Thus successful area bombing was carried out under instrument flying conditions by adjusting two Elektra stations so that two of their lanes would intersect over the chosen target.

2. *Sonne*

About 1943 a modification was made to the Elektra system to form Sonne. This consisted in linearly varying the radio frequency phases of two of the three antennas, after a starting signal. This simple addition causes each of the equisignal zones to rotate about the transmitter as a center until, after 60 seconds, each comes to occupy the initial position of its next neighbor. Thus at any point the navigator will hear a number of dots or dashes (say 22 dots) followed by an equisignal (say of 4 seconds duration) and by another number of dashes or dots (say 34 dashes). This sequence indicates that the navigator is in a zone where dots are initially heard and that he is four-tenths of the way from a dash-dot boundary to a dot-dash boundary. If he has properly identified the zone, his line of position can be found from charts that show the key orientations. Similar listening to a second Sonne station provides a fix. A great advantage of the system is that the position is found by simple listening to any ordinary receiver. The reading accuracy is good, having a minimum error of $\frac{1}{8}^\circ$/dot, but the errors of propagation mentioned above reduce the real accuracy to the order of $\frac{3}{4}$ to 1° by day and from 1 to 2° at night, with occasional errors of as much as 6 or 8°. The reliable range for stations in the 300 kc frequency region is six or eight hundred miles and useful signals are sometimes had up to two thousand miles from the stations.

3. Radar

Radar, of course, is the most discussed series of electronic devices to come out of the recent war. The techniques are too well known to need discussion here and the applications to navigation are limited, primarily by the relatively short ranges available except for observation of or from very high flying aircraft. Radar is an excellent extension of visual pilotage and some forms, such as the Ground Controlled Approach landing system, are of great advantage to the navigator. Thus harbor entrances can be made in fog, and a cross country aircraft can fly successfully from point to point over land, so long as sufficiently close attention is paid to the radar screen so that no mistakes in identification are made. Radar observation of aircraft and ships (over limited distances) from the ground solves the problem of identification of place, but replaces it with the problem of identification of the vehicle, which, in turn, can be overcome by the use of radar beacons; but need of complex networks for a complete solution to the navigational problem makes it appear that radar will be used primarily for collision prevention and that navigation will be carried out by other devices designed especially for the purpose. Such devices might be similar to or improvements upon those mentioned in the next succeeding paragraphs.

4. Gee

Gee is a powerful system, operating at frequencies between 25 and 80 Mc, that requires no transmission from the navigator's vehicle. It was the first of the pulsed hyperbolic grid-laying devices and was put into operation early in 1942. The method depends upon the measurement of the difference in the time of arrival (or, more precisely, the phase) of two trains of uniformly spaced pulses received from two well-separated points. A constant time difference identifies a line of position which, neglecting the oblate curvature of the surface of the earth, is a hyperbola with the two transmitters as foci. A similar measurement is made simultaneously upon a second pair of pulse trains radiated by one station of the first pair and a third station. The receiving equipment is simple and compact, and yields fixes with great rapidity. There is, of course, an infinite family of hyperbolas generated by each pair of stations, although the resolving power of the receiver is finite. Because the hyperbolas diverge with increasing distance from the transmitters, the accuracy of the system decreases with increasing range at a much more rapid rate than is the case with beacon responder methods. The intrinsic accuracy is high, however, and the average error of fix varies from about 200 yeards near the transmitters to about 5 miles at the maximum range, for high-flying aircraft, of 350 to 400 miles.

5. Loran

Loran, a hyperbolic system very similar to Gee, operates in the neighborhood of 2 Mc. At this relatively low frequency, transmission ranges are much greater, especially at low altitudes, than they are at Gee frequencies. The lower frequency, however, makes the use of much longer pulses essential and the measuring technique must be somewhat more complex if similar accuracy is to be attained. At this frequency, sky-wave transmission is good to a distance of 1600 miles at night, but is not effective by day. Ground-wave ranges are, typically, 800 miles over sea water in the daytime and 600 miles at night. Over land, however, the ground-wave range is so small that useful service can be provided only in special cases. Except at very low altitudes, Gee is a more useful system over land; but over sea water Loran gives comparable or greater accuracy at much greater distances, the average errors of fix ranging from 1000 feet at short distances to perhaps 1½ miles at 800 miles. At night the sky-wave service is available with average errors ranging from 1½ miles at 300 to about 8 miles at 1600.

For nighttime service only, the use of sky waves permits the exploitation of very long baselines which greatly reduce the geometrical dilution of the potential accuracy. In the case of two long baselines at right angles, the average errors of fix are small and remarkably constant, ranging from 1 to 2 miles over an area of a million square miles or more.

The use of the hyperbolic principle (or more properly, the use of long baselines) operates to reduce the pernicious effects of variations in the effective velocity of propagation of radio waves. Thus the accuracy of Loran, even for the second-rate sky-wave service, is equivalent to that of a direction-finding system with average errors of not more than 0.1 to 0.3° arc.

6. Low Frequency Loran

A Low Frequency Loran system has been tested in two areas at a frequency of 180 kc, and has given successful operational service in central and northern Canada. The characteristics of LF Loran differ greatly from those of Loran at 2 Mc, primarily because it is necessary to use pulses so long that ground waves and sky waves cannot be resolved. This factor results in large timing errors which are partially compensated by longer baselines. At this frequency the atmospheric noise level varies greatly with time and with geographical position, giving rise to large differences in useful range. In spite of these limitations the average errors of fix are between 5 and 20 miles out to the nominal maximum range of 1000 miles over land or 1500 miles over sea. The system is particularly useful in the high latitudes where the noise level is low.

One variant of LF Loran, called "cycle-matching," is of particular interest, in spite of limited testing. In this embodiment of the hyperbolie principle the pulse envelope is used to identify a particular radio frequency cycle, and the cycle itself is used for a precise measurement. In a brief trial of this method the correct cycle was identified in three-fourths of the attempts and the average deviation of the readings within the cycle was 0.15 microsecond. This figure corresponded to an average error of the line of position of 160 feet at a distance of 750 miles, Although this is almost the extreme range at which this method can be used, because of sky-wave interference with the necessary ground waves, and although the problems of cycle identification remain to be solved, the technique seems worthy of further exploration.

7. *Decca*

Finally we find a very interesting continuous-wave system, Decca, which has given some service at about 100 kc. Decca is a hyperbolic system which detects and measures the relative phase of the carrier frequency cycles. It therefore has very high resolution for a system at these frequencies, changes in position of a few yards being detectable. Unfortunately, the range of the system is severely limited by sky-wave transmission which appears in random phase and is of great amplitude at distances of a few hundred miles. Thus the precision is high at short ranges but the readings become ambiguous beyond 150 or 200 miles. Experiments with this system are being conducted, we understand, at frequencies of 10–15 kc. In this case the satisfactory range should be greater, because of the increased efficiency of ground-wave transmission, and the limiting accuracy should be reduced in proportion to the frequency.

IV. POSTWAR PROPOSALS

All of the systems mentioned in part III were put into operation and used to good advantage during the war. They are all still in use to a greater or lesser extent and some of them have a good chance of continued service for many years.

The new techniques available since the war may be used to aid navigation in many ways, and many proposals have been made. To some extent the very number of proposals has militated against careful consideration, to say nothing of exploitation, of any of them. One or two examples to be cited presently will suffice for the present discussion.

1. *Omnirange*

There is one system, the omnidirectional beacon or omnirange, however, that was not used during the war but that has the advantage of a number of years of study and experimentation in the hands of the Civil Aeronautics Authority. It is of particular importance because it will become the standard device that replaces the radio range along the airways of the United States. It is perhaps unfortunate that the accuracy and radius of operation are not significantly greater than those of the radio range, but the system has the merit of simple geometrical relationships that will permit the use of various simple computers to aid automatic flight.

The omnirange, which operates in the very high frequency region, is one of the azimuth-determining systems. The transmitting antenna radiates a cardioid pattern of field intensity that rotates in azimuth at a low audio frequency. The amplitude of the signal received at a fixed point therefore varies sinusoidally with an absolute phase that depends upon the azimuth of the receiver as seen from the transmitter. This phase is continuously and automatically measured, with reference to a second modulated signal radiated as a frequency modulation from the same transmitter, to give a direct indication of the azimuth. The average accuracy of the system is reported to be about 3° and the range, at low-flying altitudes, is at least 50 miles. The omnirange ordinarily shares its site with a responder beacon that can be interrogated by airborne pulse interrogators to measure the distance to the beacon. The combination then determines position in polar coordinates, although unfortunately the precision of the measurement of azimuth does not compare with the precision of measurement of the radial distance. Since the geometrical relations, as mentioned above, are simple, it is possible to build a computer to operate from the output of the omnirange and distance-measuring receivers that will continuously indicate to the pilot both the deviations from any chosen course (anywhere within the range of the beacon) and the distance to go to reach the objective at the terminus of the chosen course.

2. *Teleran*

As an example of the more complex networks of navigational aids that may come into existence, it may be sufficient to mention Teleran.

This is a proposal to use a network of ground-based radar stations (with airborne transponder beacons for aircraft identification) to keep continuous track of all aircraft in the air over a large area. This information, or rather the parts of it of interest to a particular aircraft, is then added to a map of the immediate vicinity with notes about the weather, the state of airport runways, etc., and transmitted to the aircraft by television. In this way the indicator oscilloscope of each aircraft navigator continuously conveys the information he needs—data on winds and on other aircraft while he is in transit and information as to landing conditions as he approaches his destination. A large network of ground equipment is, of course, necessary, but there is a great advantage in placing the complex equipment on the ground where size and weight are not of major importance and where adequate supervision and maintenance are available. The use of Teleran, or similarly complex networks, depends upon careful study of the economic gains and gains in safety to be derived from the careful control of aircraft in flight.

V. Considerations of Range and Accuracy

It may seem, at first glance, that the characteristics of these several systems are almost unrelated, but actually the ranges and accuracies fall into simple patterns. Nothing, in this section, can be said about cost or operating convenience, but it is possible to place fairly definite limits upon what navigational systems can and cannot do.

1. Classification of Systems

There are three kinds of systems which we may classify according to the shapes of the position lines they generate; *radial,* for the various forms of direction-finding; *circular,* for the distance-measuring responder systems; and *hyperbolic,* for the time-difference methods.

With the exceptions of High Frequency Direction Finding, a sky-wave system, and Decca, which is a ground-wave system suffering from sky-wave interference, the useful ranges of the various devices increase with wavelength. Under the best conditions, in fact, the maximum range for ground (or space) waves increases nearly as fast as the fourth root of the wavelength. The average errors of the various systems, in general, increase in proportion to the wavelength except for discontinuities produced by the changes from one technique to another. We must now explore these relations in some detail.

2. Ground- and Sky-wave Range

The chief factors that control the useful range of a radio signal are the shape and electrical conductivity of the earth, the state of ionization and noise-producing electrification of the atmosphere, the bandwidth and, particularly, the frequency of the signal itself. Consider first the ground wave, or space-propagated wave, that travels along or near the surface of the earth. Let us assume a reasonably high power and what we may call a normal bandwidth that depends on frequency. We may further presuppose a fairly high level of natural noise which increases, in general, with the wavelength. Under these limitations we can draw the diagram of Fig. 1, the data for which are the satisfactory ranges of navigation systems themselves, as indicated by the names of the systems. The diagram is almost self-explanatory. At the lowest wavelengths, transmission ranges are essentially optical, depending primarily upon the elevation of the receiver. Through the medium wavelengths the range over sea water depends very little upon altitude, but over land the range is reduced and is sensitive to elevation. At the longest wavelengths the range is great and depends almost entirely upon wavelength.

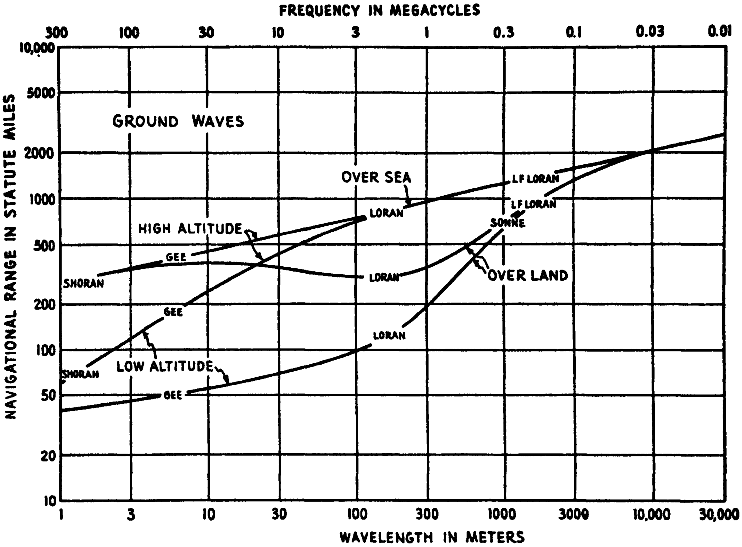


FIG. 1.—Maximum ground-wave transmission ranges as a function of wavelength. Power of the order of 100 kw and a high noise level are assumed.

To the ranges of Fig. 1 we must add the sky-wave factors shown in Fig. 2. These have been drawn neglecting seasonal variations, for the sake of simplicity, and in general the limitations are the same as for Fig. 1. Very great ranges may be had at frequencies of the order of 10 Mc or 30 kc, but service is available to only a few miles by day and to less than 2000 miles at night in the neighborhood of 1 Mc.

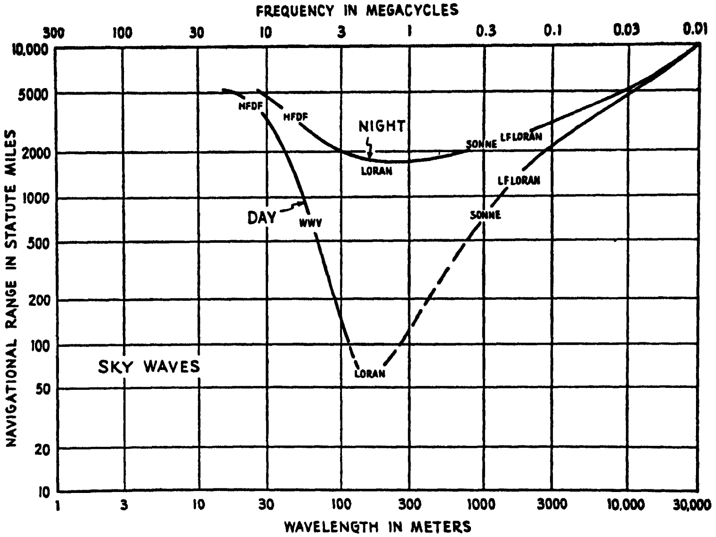


FIG. 2.—Average sky-wave transmission ranges comparable to those of Fig. 1.

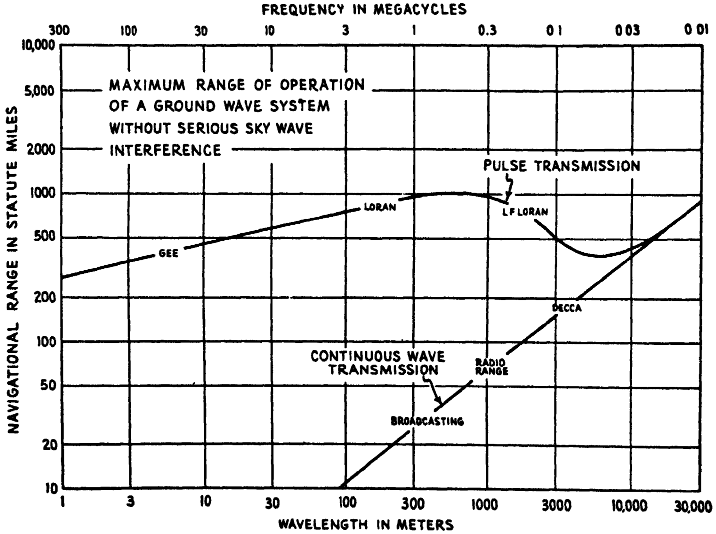


FIG. 3.—Maximum ranges at which ground-wave transmission is essentially free from sky-wave interference.

The strong sky-wave transmission at wavelengths greater than 1000 or 2000 meters invalidates the longer ground-wave ranges shown in Fig. 1. This effect is shown in Fig. 3. Here the explanation is simplest if we start with the straight line at the right. A continuous-wave system will take advantage of all the characteristics of ground-wave transmission at short distances from the transmitters, where the ground wave is far more intense than the reflected sky waves. A few miles or a few hundred miles away, the ground-wave field intensity, which is decreasing with increasing distance much faster than are the sky waves, falls to such a level that sky-wave interference begins to control the characteristics of the signal. Since the phase of a sky wave, relative to that of the ground wave, is random, the resultant signal fades. If the navigation system is one that measures radio-frequency phase, such as Decca or cycle-matching LF Loran, the indicated phase is random and the readings bear little relation to the position of the navigator. The straight line has been drawn to indicate the maximum useful range for "pure" ground waves.

3. Range of Pulse Systems

Pulse systems in the high and medium frequencies have a great advantage in this respect because, by use of short pulses, the ground wave may be resolved and observed without sky-wave interference. Thus the full transmission range is available for ground-wave systems. Below about 1000 kc, however, it is difficult to radiate a pulse sufficiently short to be received completely free of overlapping sky waves. The front part of the pulse is still "clean" and may be used, but with increasing wavelength the amplitude of the uncontaminated portion decreases (because the pulse length increases and the slope of the leading edge decreases); thus the effective range decreases. In the neighborhood of 200 kc this effect becomes accentuated because the practical limit of physical height of antenna structures is reached. Below this frequency the antenna Q rises sharply and the pulse length increases very rapidly. At very low frequencies, "pulse" transmission cannot be distinguished from continuous-wave transmission. This diagram is of great importance because it shows why systems with the high accuracy that can be attained through ground-wave transmission cannot be expected to operate at ranges of more than a thousand miles.

The data of Figs. 1, 2, and 3 are combined in Fig. 4, from which we can estimate the types of transmission and useful range that apply at any wavelength. The daytime sky-wave curves and the curves for ground-wave transmission at the longer wavelengths are dotted in the regions where they are of only academic importance. The shaded areas, for ground-wave transmission, embrace the variation of range with altitude of the receiver. The limits are roughly 100 and 30,000 feet. The ranges shown apply approximately for any kind of system, although the narrow-band systems will have somewhat greater range than the pulse systems, for the same peak power. The pulse systems are likely to have the greater range if average radiated power is taken as the criterion. All the curves have been drawn conservatively to indicate ranges that should be obtained a high percentage of the time.

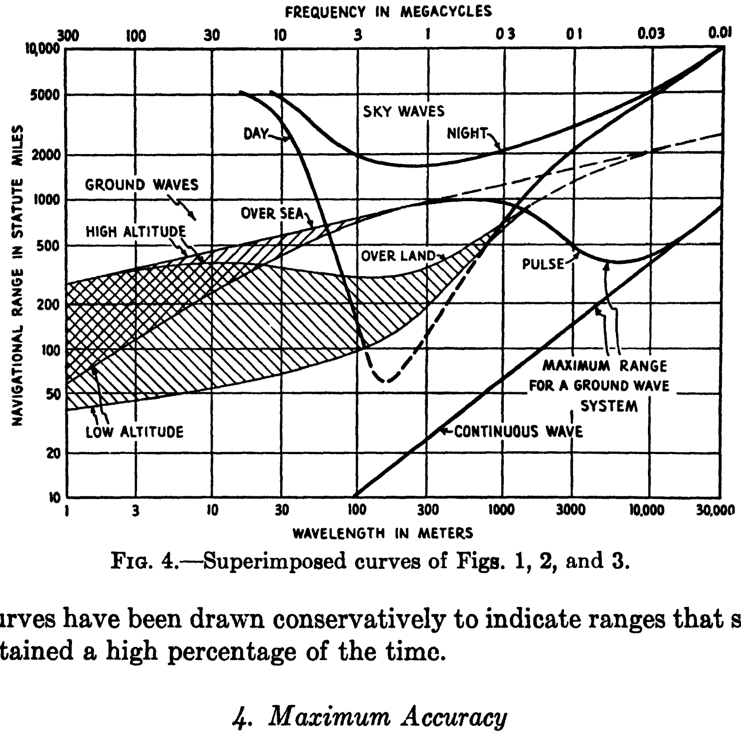


FIG. 4.—Superimposed curves of Figs. 1, 2, and 3.

4. Maximum Accuracy

If we neglect, for the moment, direction-finding methods we can form some useful conclusions about the potential accuracy of pulse and phase systems by considering the effect of changing wavelength. Let us consider the pulse systems first. The effect of the Q of electrical circuits and antennas, as well as the increased noise associated with wide-band receivers and the interference of radio signals with other services, is to limit the steepness and shortness of usable pulses. In practice it becomes very difficult to generate and efficiently radiate a pulse whose length is less than about 50 cycles of the carrier frequency. This figure does not appear to vary appreciably with frequency except when the antenna Q, mentioned above, begins to increase very rapidly with wavelength. Now if a time measurement is to be made from one pulse to another, as by having both pulses displayed on a radar trace, the measurement can be made to about one-fifth of the length of the pulses. The measurement may therefore be expected to be accurate to about 10 periods, in time, or 10 wavelengths, in space. The "least reading" in distance is taken as 5 wavelengths since a geometrical factor of 2 is always gained under optimum conditions. For example, if the distance of a radar target increases by 5 wavelengths, the time interval from transmission to reception of an echo increases by 10 periods. If two such measurements are made at right angles to each other, we can estimate the average error of fix under this optimum geometrical condition. This estimate has been made on the conservative assumption that the average error of fix will be 2 $\sqrt{2}$ times the least reading, where the $\sqrt{2}$ is the factor derived from the combination of two uncorrelated measurements at right angles, and the 2 is intended to cover the discrepancy between "laboratory" and "field" accuracies.

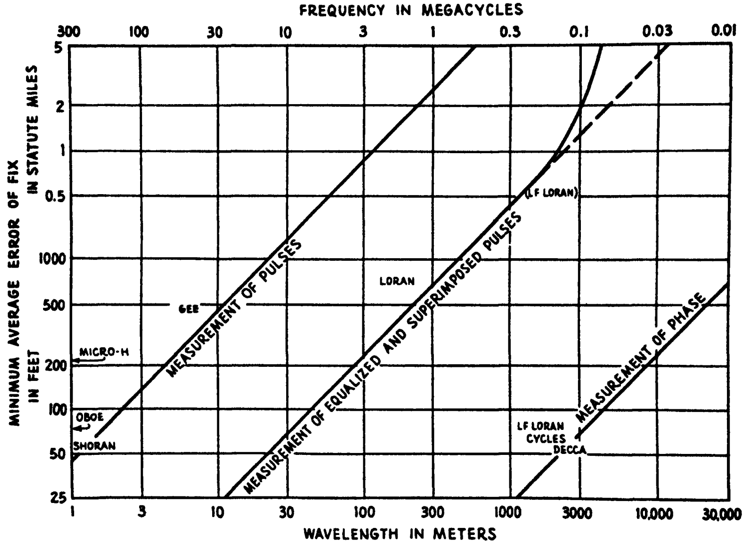


FIG. 5.—The minimum errors of fix attainable under the best geometrical conditions with no transmission errors by three standard measuring techniques.

Fig. 5 shows, at the left, a line drawn according to these constants. This line shows that the minimum average error of fix, or the average error of fix under the best geometrical conditions, varies linearly between about 50 feet at a wavelength of 1 meter to 5 miles at a wavelength of 600 meters.*

---

* Data are given on the figures for Shoran, a system that has not been discussed here because it is primarily useful for blind bombing and surveying rather than general navigation. It is a system consisting of an airborne pulse interrogator with

An improvement in accuracy can be realized by a pulse-matching technique used, to date, only in the case of Loran. If two pulses to be compared are carefully made similar when they are radiated, and have their amplitudes made equal in the receiver, the pulses can be visually superimposed and compared with an accuracy of 1% of the pulse length. This makes possible an enhancement of 20 times in the accuracy, an improvement shown in the center curve of Fig. 5. This curve departs from linearity at wavelengths above 1500 meters because of the effects of antenna *Q* mentioned above. At very long wavelengths the errors for pulse transmission would rise as the fourth power of the wavelength.

The next, and, so far as we know, final step in the direction of greater accuracy is the measurement of radio-frequency phase. Depending upon the technique, the reading accuracy may be expected to lie between 1% of a cycle and 1° of phase: The third line on Fig. 5 is drawn assuming a precision of $\frac{1}{200}$ wavelength, or 100 times the Loran reading accuracy. This method, as explained above, is useful only for ground-wave transmission. The one exception to this statement is in the case of phase measurement of a modulated wave, in which case the modulation wavelength must be considered as the criterion. In this embodiment the wavelength in Fig. 5 must be taken as the wavelength of the modulation envelope with a consequent decrease in precision.

The three curves of Fig. 5 embrace all of the techniques so far developed. Other orders of precision are, of course, possible but it is difficult to think of new mechanisms that do not fall into one of the classes shown. The names of a number of systems have been inserted in Fig. 5 at points indicating the appropriate wavelengths and accuracies. In the case of each system with average errors greater than those indicated by the appropriate line, it is easy to find a modest improvement in technique that would reduce the average errors. LF Loran, using ground-wave transmission, is the only case in which the potential accuracy is greater than that indicated by our criterion. As one of those most closely associated with the development of this system, the writer can testify that it required extreme effort to produce pulses sufficiently short to achieve this result, and that this labor was hardly justified by any great operational advantage.

---

distance-measuring equipment and two ground-based responder beacons, and its actual accuracy has not yet been exceeded. This achievement stems primarily from the careful design of the system for its intended function. The choice of constants, such as pulse length and bandwidth, is admirable. The average errors are primarily propagational and are of the order of 20 yards at distances up to about 250 miles.

Micro-H is a similar system using an airborne radar set for the same purpose. It is now obsolete.

A study of the accuracies of the various systems indicates exactly what we might expect; that it is relatively easy to approach the criteria indicated in Fig. 5 but difficult to exceed the indicated accuracies by a worth while margin.

Fig. 5 has given us the reasonable minimum errors of measurement. To these we may add the errors inherent in E-layer sky-wave transmission, to deduce the minimum errors of received sky-wave pulses. These curves are shown in Fig. 6 and in the composite ground- and sky-wave pattern of Fig. 7. The derivation of these curves is purely empirical but they are believed to be reasonably accurate. There is no corresponding curve for phase measurement because, as mentioned above, sky-wave transmission invalidates phase measurement except in the case of modulated waves, and in this case most of the errors would be beyond the limits of Figs. 6 and 7. Similarly, there are no curves for F-layer transmission which would multiply the sky-wave errors shown by a factor of the order of 10.

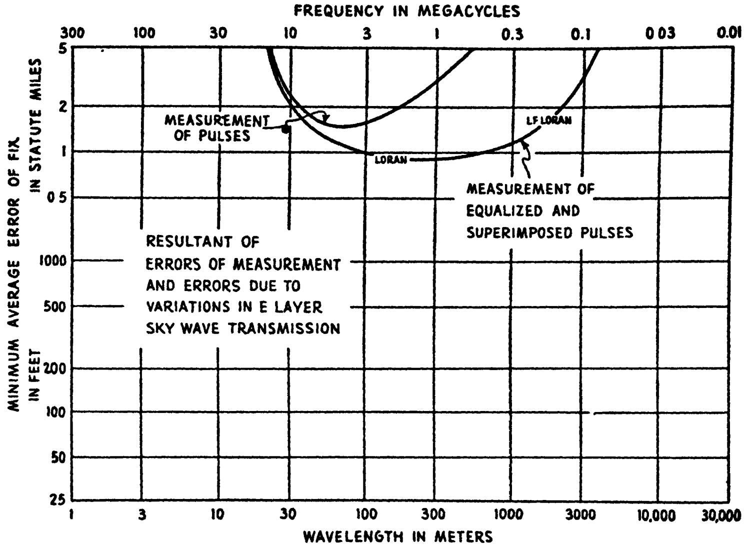


FIG. 6.—Minimum errors, comparable to those of Fig. 5, but including E-layer sky-wave transmission errors.

5. *Geometric Factors*

From a study of Figs. 4 and 7, now, we can deduce the range and maximum accuracy of any radio aid to navigation with the exception of direction-finding. So far, however, we have the accuracy available at only one special point in the service area of a system. Any departure from that point will lead to increased errors. We must now examine the geometrical factors that operate to increase the errors of any practical system. To do this we shall return to the classification of circular, hyperbolic, and radial systems. For the first two the minimum average errors of fix can be estimated from Fig. 7. Radial systems will be discussed after dealing with the other two.

Figs. 8, 9, 10, and 12 show the rate of increase of the average error of fix with distance, when distance is measured in the direction shown in each of the small inset diagrams. The unit of distance is the length of the baseline between the transmitters used to obtain a fix, as shown by the dotted lines in the key diagrams. The unit of average error of fix for Figs. 8, 9, and 10 is the minimum average error of fix obtained from Fig. 7.

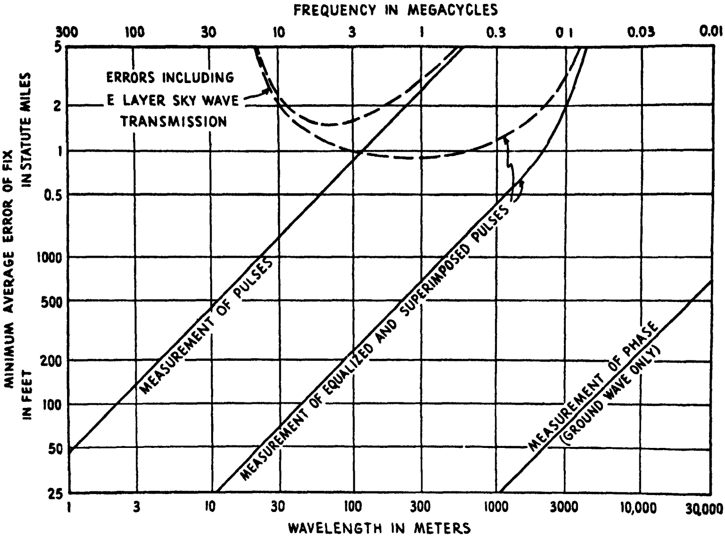


Fig. 7.—Superimposed curves of Figs. 5 and 6.

In any direction other than those shown, the errors increase more rapidly. This effect may be estimated quite well because the accuracy of a system is, in general, a simple function of the angle subtended by the stations as seen from the navigator's position. Therefore the error at any point is about the same as the error at a point in the best direction (shown on Figs. 8, 9, and 12) where an equal angle is subtended by the transmitters. In the cases of Fig. 10 and the three-station system of Fig. 12, the errors do not vary appreciably with direction.

Fig. 8 represents the simplest case, that of a circular, or distance-measuring system. Here the error of range measurement does not depend on distance so that variations of the average error of fix depend only upon the crossing angle of the two circular lines of position. The error is infinite at zero distance, because the lines of position are tangent there; it falls to unity when the distance is half the baseline and the crossing angle is 90 degrees. Thereafter the errors increase almost in proportion to the distance. At three times the length of the baseline, for instance, the average error of fix is 3.1 times the minimum average error of fix.

Hyperbolic systems give more complex patterns, some of which are shown in Fig. 9. The simplest and worst orientation of stations is the triplet in which two baselines are laid end to end as shown in the upper part of the small diagram. Here the errors increase nearly as the square of the distance because the hyperbolas diverge while the crossing angle decreases, but fortunately the inherent minimum average errors are small so that useful service has been provided in areas where the "dilution factor" was as high as 50. It should be noted that there is no point in the neighborhood of a triplet where the average error of fix is less than about 1.3 times the minimum average error.

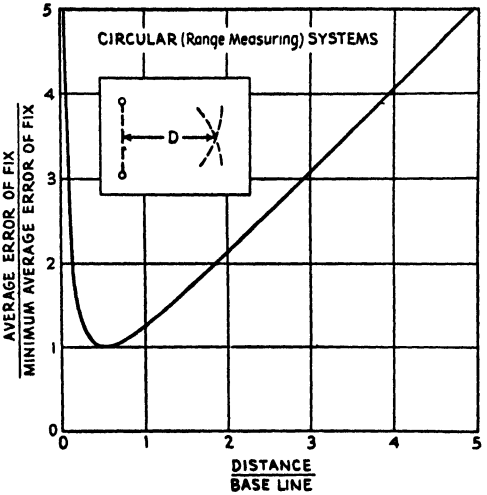


FIG. 8.—The relative errors of fix of circular systems as a function of distance.

For a given length of baseline the best orientation of stations is that which gives a crossing angle of 90 degrees at the navigator's position. This condition is assumed for the curves marked "Two Pairs" on Fig. 9. Thus the general solution to any hyperbolic problem will lie between the solid curves of that figure. So far as we know, the same construction applies for the case of unresolved sky-wave transmission, where the timing errors do not vary greatly with distance. Resolved sky waves, on the other hand, yield smaller timing errors at longer distances so that the increasing geometrical errors are partially cancelled. The corresponding curves, for E-layer transmission, at medium frequencies, are shown dashed in Fig. 9. The unity error to be used with these dashed curves is, of course, the sky-wave value from Fig. 7.

The important special case of two pairs of hyperbolic stations with a common center and baselines at right angles (the hyperbolic quadrilateral) is shown in Fig. 10. This is the best possible orientation of stations when navigation is desired only within the confines of the square. Outside the square, as shown by the error curve, the crossing angles degenerate very rapidly, but within the square the errors are small and sensibly constant.

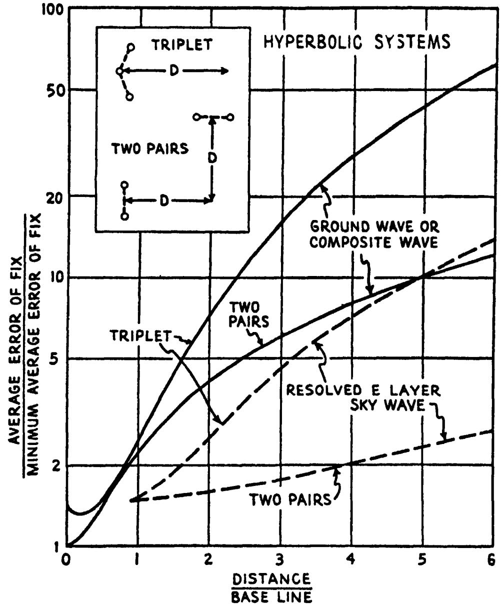


Fig. 9.—The relative errors of fix of hyperbolic systems as a function of distance.

The treatment of Figs. 5 through 10 has summarized the accuracy of circular and hyperbolic systems. Radial, or direction-finding, systems may be compared to the others by a correspondingly simple treatment. The standard error in this case is the average error of the line of position at a distance equal to the length of the baseline. This LOP error may be estimated from Fig. 11 in which lines are given for average angular errors of 1, 2.0, and 5 degrees. Nearly all direction-finding techniques seem to give average errors near one of these values; some examples are

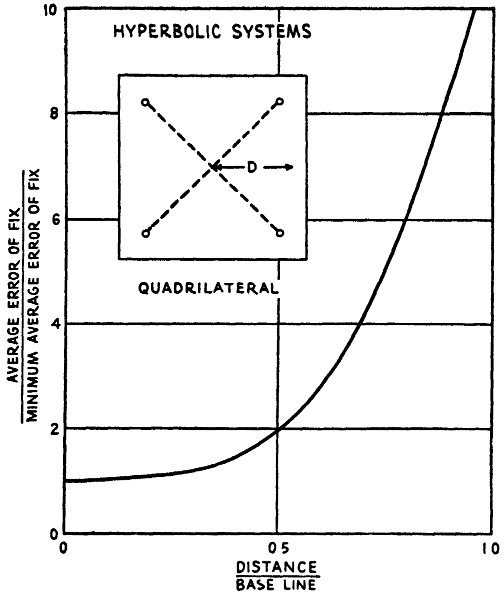


FIG. 10.—The relative errors of fix for the special case of the hyperbolic quadrilateral.

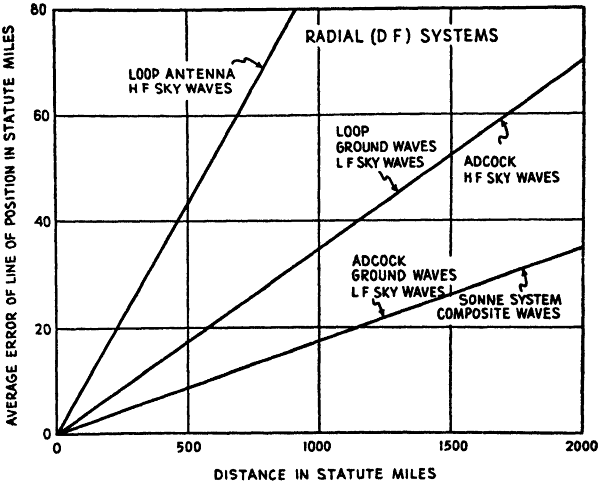


FIG. 11.—Line of position errors of a number of direction-finding methods as a function of distance.*

* The Adcock antenna is a design that is particularly free from the pernicious effects of varying polarization of sky-wave signals. Its use therefore results in reduced errors when receiving sky-wave signals. The antenna is inherently large so that it cannot be used on aircraft and is not often found on shipboard.

noted on the figure. The average error of fix at any distance may be found from Fig. 12, which is similar to the other figures just cited except that the unity error, in this case, is for a line of position rather than for a fix. The minimum error of fix at any point is not unity but has a value of 0.92.

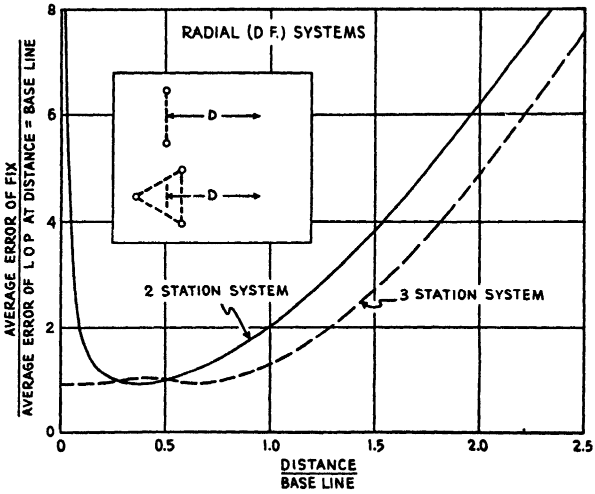


FIG. 12.—The relative errors of fix of radial systems as a function of distance.

6. *Range vs. Accuracy*

Systems using a combination of methods, such as the omnidirectional range with distance-measuring equipment, must be examined individually. In this case, the distance-measuring equipment ordinarily has much smaller linear errors than the omnidirectional range. The average error of fix is therefore practically equal to the line of position error of the omnidirectional range.

The treatment described in the last few pages will permit the estimation of the range and accuracy of a proposed radio aid to navigation. The data from Figs. 4 and 7 may be combined into many useful patterns provided it is remembered that the practical errors are always greater than the minimum average error, and are, in many cases, from three to ten or more times as great.

Perhaps the most useful of these secondary diagrams is shown in Fig. 13. By intercomparison of Figs. 4, 7, and 11, we can determine the maximum accuracy of the various possible techniques that can be used at the wavelengths requisite for the achievement of any desired range. A number of these lines showing the potential behavior of some of the better systems are given in Fig. 13. In each case the baseline is always assumed to be equal to the range, for a shorter baseline would result in increased errors. Thus the diagram shows errors that cannot be reduced except by improvement in measuring technique, for ranges less than 1000 miles, or by reduction of the variations due to sky-wave transmission for the longer ranges.

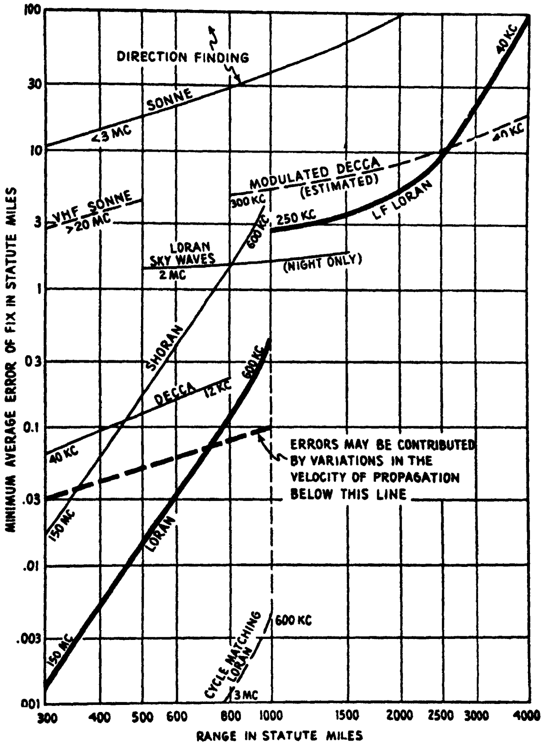


FIG. 13.—The minimum errors of fix attainable by various techniques as a function of maximum required range. The heavy lines represent the best techniques so far reduced to practice.

The heavy line of Fig. 13 indicates the minimum errors of fix attainable at any range from 300 to 4000 miles. At 1000 to 1500 miles, for instance, the line marked "Standard Loran Sky Waves" is based upon the actual fluctuations in the time of arrival of sky waves. This level of error (1¼ miles) is not adopted for the heavy line because operation of this sort can be had only at night. The LF Loran curve, at the same range, shows about twice the minimum errors because ground waves interfere with the sky-wave transmission, which would be more accurate if it could be observed alone. The curve called "Modulated Decca" is drawn for a mythical system on the assumption justified by estimates of the sky-wave delay time at low frequencies, that, by the use of the LF Loran pulses, the composite-wave timing errors are reduced to about one-half of their normal cw value. At very long wavelengths, or very long ranges, the modulated Decca technique (or any other phase system operating with a modulation frequency low enough to avoid problems of ambiguity) would realize greater accuracy than LF Loran, which would be striving for pulse matches with absurdly long pulses. The actual errors to be expected in the case of "Modulated Decca" are probably uncertain by a factor of 2. The curve is dashed because this system is not operational.

For ranges less than a thousand miles, phase measurement on ground-wave transmissions is capable of yielding average reading errors of the order of $\frac{1}{1000}$ of a mile or less, depending on the range. There is probably little practical use to be expected from this technique because the velocity of radio waves in the atmosphere is not sufficiently constant to justify the effort. The dashed heavy line on Fig. 13 is drawn for errors equal to 1/10,000 of the range. Exploitation of systems having errors smaller than this will require the determination of the velocity of propagation for the altitude and climatic conditions in question. Probably errors 10 times below this level can be achieved by careful consideration of this point.

The merits or potential merits of a navigation system may easily be determined by the methods of this paper because the three basic factors that determine range and accuracy have been treated individually. Figs. 1 to 4 and the sky-wave curves of Figs. 6 and 7 summarize the effects of radio transmission; Figs. 5 and 7 exhibit the inherent errors that are controlled by the choice of technique and the limitations of electronic equipment; and Figs. 8, 9, 10, and 12 summarize the geometrical relationships in the various systems.

If greater detail should be desired, families of curves, instead of single lines, may be drawn to show the effects, for instance, of variations in transmitter power, noise level, bandwidth, season of the year, and time of day. Unfortunately many of the data are lacking for the immediate construction of such diagrams. Even in their absence, however, it is probable that the figures shown herein give solutions within 20% in range and within a factor of 2 in accuracy for most practical situations. Additional data, in some cases, would improve the precision of the present curves.

Two fundamentally important conclusions are to be derived from Fig. 13 and from this whole discussion:

a. At distances less than 800 or 1000 miles, accuracies of 0.1 mile or better are easily attainable;

b. At distances greater than 1000 miles, the *minimum* average errors of fix are of the order of 3 to 5 thousandths of the distance, and the practical average errors over a large area may be as much as 1% of the range.

Author Index

Numbers in parentheses are reference numbers. They are included to assist in locating references in which the authors' names are not mentioned in the text.

Example: Allison, H. W., 58 (88), *64* (88), indicates that this author's article is reference 88 on p. 58. Numbers in italic refer to the page on which the reference is listed in the bibliography at the end of each chapter.

A

Abbot, T. A., 220 (11), 239, 264 (11), *266* (11)
Abramov, A., *125* (130)
Afanasjeva, A. V., 73, 108 (175), *121* (39, 40), *122* (62), *126* (175), *129* (232)
Agar, W. O., *345* (23)
Ahearn, A. J., 71, 78, *120* (20)
Alford, A., 376, *380* (15)
Allen, E. W., 405 (17), 417, 419, 420 (17, 25), 420 (24), 421, *422* (17), *423* (22, 24, 25)
Allen, J. S., *124* (113)
Allison, H. W., 58 (88), *64* (88)
Alvarez, L. W., 276 (10), 311, *315* (10), *316* (30)
Antonov, V. A., *125* (141)
Appleton, E. V., *344*
Aram, N. W., 415, *422* (21)
Aranovich, P. M., 108 (176), 116, *126* (176)
Aranovich, R. M., *129* (233)
Armstead, F. C., 265 (128), *268* (128)
Arnold, H. D., 24 (47), *63* (47)
Aston, F. W., 220, 228, 238, 240, 250, 256, *265* (2), *266* (26), *267* (46, 48, 60, 87), *268* (100)

B

Baarden, J., 50, *64* (82)
Bainbridge, K. T., 224, 226, 227, 238, 242, 244, 245, 246, 247, 254, 256, *266* (16), *267* (50), *268* (93)
Baldwin, M. W., 165, *166* (21)
Ballard, R. C., 141, *165* (1)
Barnes, D. E., *316* (19)
Bartels, J., *344* (4)
Bartky, W., 226, 247, *266* (20)
Bateman, R., *344*
Bates, D. R., *344*
Bauchman, R. W., 419, *423* (28)
Bauer, N., 254 (92), *268* (92)
Bay, Z., *123* (85), *127* (187)
Beach, J. Y., 254 (92), *268* (92)
Becker, A., 79 (4), 82, 83, 88, *120* (3, 4, 13)
Becker, G. A., *130* (245)
Becker, J. A., 4, 61, *62* (6, 9), *64* (89), 70, 73, *121* (32)
Bekow, G., 74, *127* (188)
Bendt-Nielsen, B., 228 (32), *266* (32)
Benjamin, M., 9, 10 (21), 17 (34), *62* (20, 21), *63* (34)
Beringer, R., 376, 377 (20), *380* (20)
Berkner, L. V., *344*
Berlin, T. H., 295, *316* (21, 25)
Berry, C. E., 220 (7), 260 (7), 261 (114), *266* (7), *268* (114)
Bethe, H., 87, 91, 92, 105, *120* (16), *127* (189)
Bey, A., *122* (68)
Beyer, O., *122* (71)
Beynon, W. J. G., *344* (6, 13)
Bhawalker, P. R., 90 (63), *122* (63)
Binnian, W., 419, *423* (28)
Blackwell, H. R., 146, *166* (7)
Bleakney, W., 208, 211 (20), *218* (20), 220 (9), 231, 238, 239, 242, 247, 250, 251, 254, 260 (9), 265 (9), *266* (9, 35), *267* (49, 59, 63, 83)
Blewett, J. P., 1, 2, 14, 30 (1), 32, 51, 52, 55, 61, *62* (1, 3) *64* (84), 288, *316* (15)
Bohm, D., 295, *316* (22)
Bojinesco, A., *124* (114, 115)

Bolton, J. G., 356
Bondy, H., 238, 246, *267* (55)
Booker, H. G., *344*, 397, 402 (10), *422* (10)
Booth, E. T., 258 (108), *268* (108)
Borzyak, P., 115, *124* (116, 117)
Bowie, R. M., 199 (14), *218* (14)
Bredennikowa, T. P., 21, *63* (40)
Brinsmade, J. B., 82, 83, *120* (9)
Brobeck, W. M., 276 (10), 301 (18), *315* (10), *316* (18)
Bronstein, 82, *129* (224)
Brown, B. B., 37, *64* (72)
Brown, H., 239, 265 (122, 123), *267* (67), *268* (122, 123)
Brown, J. B., *122* (64)
Brüche, E., 197, *218* (11)
Bruining, H., 66, 68 (66, 86, 87, 120, 216), 70, 72, 76, 77 (65, 86), 78 (86, 118), 84, 88, 90, 98 (87, 120), 99 (119), 100, 105, 106, 107, 108, 110, 116, 117, *121* (41) *122* (66), *123* (86–89), *124* (118–120), *127* (190) *128* (216)
Brunner, W., 384 (1), *421* (1)
Buckingham, R. A., *344* (10)
Buechner, W. W., 271 (2), *315* (2)
Burgers, W. G., 9, 11, *62* (19), 77 (65), *122* (65)
Burgess, J. S., 68 (250), *130* (250)
Burrill, E. A., 271 (3), *315* (3)
Burton, J. A., 18, *63* (37)

C

Carnahan, C. W., 415, *422* (21)
Cartan, L., 226, *266* (22)
Cazalas, A., 176, *218* (5)
Chamanlal, C., *344*
Chamberlin, O., 265 (130), *268* (130)
Champieux, R., 36, 38 (69), 40, *63* (69)
Chapin, E. W., 392 (9), *422* (9)
Chapman, S., *344* (4, 16)
Charlton, E. E., 281 (12), *315* (12)
Chaudri, R. M., 71, 72, *127* (191)
Clarke, I. G., 376, *380* (15)
Classen, E. F., 415, *422* (21)
Coates, W. M., 306, *316* (27)
Cobb, P. W., 146, *166* (5)
Cockcroft, J. D., 271 (1), *315* (1)
Coggeshall, N. D., 208, 210, 211 (21, 26) *218* (21, 26), 222 (13), 239, 255, *266* (13), *267* (69), *268* (94, 95)
Cohen, A., 250, 251, *267* (84)
Condon, E. U., 256, *268* (99)
Connor, J. P., 146, *166* (6)
Cooksey, D., 276 (10), *315* (10)
Coomes, E. A., 12 (27), 32, 38 (27), 53 (27), 54, 55 (27), *63* (27), 68 (121), 71, 72, 73, *124* (121)
Copeland, P. L., 70, 71, 89 (155), 107, *121* (26), *123* (90), *124* (122), *126* (155), *129* (227)
Corson, D. R., 276 (10), *315* (10)
Costa, J. L., 238, *267* (47)
Cottony, H. V., 364, *380* (27)
Cowling, T. G., *344*
Crane, H. R., 298 (24), *316* (24)
Crichlow, W. Q., 364, *380* (27)
Crowther, B. M., 199 (15), 200, *218* (15), 238, 247, *267* (54)

D

Daene, H., 79 (14), *120* (14)
Darbyshire, J. A., 5, 11, *62* (12, 25)
Darwin, C., *345*
Davison, W. L., 220 (3), *266* (3)
Davisson, C. J., 85, *120* (10)
de Boer, J. H., 5, *62* (11), 68 (66, 87, 120), 70, 72, 77 (65), 78 (29, 118), 98 (87, 120), 99 (119), 107, 110, 117, *121* (29), *122* (66), *123* (87), *124* (118–120)
Dechend, H. V., 250, *267* (79)
Dellinger, J. H., *345*
de Lussannet de la Sablonière, C. J., 95 *121* (25)
Dempster, A. J., 224, 226, 228, 230, 231, 237, 238, 241, 243, 246, 247, 254, 256, 257, 260 (111, 112), 263, *266* (18, 20, 21, 28, 29), *268* (111, 112, 116)
Dennison, D. M., 208, 217, *218* (25), 295, *316* (21, 25)
De Vries, H., 144, *166* (3)
Dicke, R. H., 376, 377 (20), *380* (20)
Dillinger, J. R., 43, 44, 54, *64* (78, 86)
Dobroljubski, A. N., 115, *121* (42–44), *122* (67)
Dole, M., 256, *268* (96)

Dunning, J. R., 258 (108), *268* (108)
Dyatlovitskaya, B. I., *126* (166)

E

Eckersley, T. L., *345*
Edlefsen, N. E., 271 (6), *315* (6)
Eggleston, F., 384 (5), *422* (5)
Eisenstein, A., 11 (26), 12 (30), 15, 25 (52, 55), 26 (33, 58) 27 (33), 42 (51), 43, 53 (33), *62* (26), *63* (30, 33, 52, 55, 58)
Elder, F. R., 296 (20), *316* (20)
Elinson, M. I., 119 (240), *129* (240)
Elterton, G. C., 265 (120), *268* (120)
Epstein, P. S., 405, 407 (15), 419 (15), *422* (15)
Eskola, P., 25 (54), *63* (54)
Evans, M. W., 254, *268* (92)
Evans, W. E., *345* (31)
Everhart, E., 309 (29), *316* (29)
Ewles, J., 23, *63* (46)

F

Falloon, S., *345* (23)
Fan, H. Y., 36, 37, *63* (68)
Farmer, F. T., *345* (22, 23)
Farnsworth, H. E., 68 (5), 79 (5), 85, 89, *120* (5, 21), *121* (45), *124* (110)
Farnsworth, P. T., 114, *121* (27), 150, *166* (14)
Feenberg, E., 253, *268* (90)
Feldman, 349, *380* (6)
Ference, M., *346* (44)
Ferraro, V. C. A., *345*
Fineman, A., 25, 42 (51), 43, 53, *63* (52), *64* (85)
Fleming, J. A., *344* (3)
Fleming-Williams, B. D., 197, 198 (12), *218* (12)
Flory, L. E., 151, *166* (16)
Foldy, L., 295, *316* (22)
Forrester, A. T., 239, 250, 252, *267* (75)
Fowler, R. D., 239, *267* (67)
Frank, J., 256, *268* (98)
Frank, N. H., 295, 297 (23), *316* (23)
Frerichs, R. V., 111 (123), *124* (123)
Fricke, H., 117 (24), *121* (24)
Friedheim, J., 116, *127* (192)
Friis, H. T., 349, 358 (12), *380* (6, 12)
Frimer, A. I., 108, *126* (156), *130* (255)
Fröhlich, H., 91, *120* (22)
Frumin, M. I., 82, *127* (19)

G

Gaertner, H., 11, *62* (24)
Ganoung, R. E., 146, *166* (6)
Gardner, B. C., 150, *166* (15)
Germer, L. H., 85, *120* (10)
Geyer, K., 98 (159), 113, *126* (159), *128* (211)
Geyer, K. H., 70, 71, 98 (218), 104, 105, 111, *128* (218), *129* (225)
Gille, G., 116, *127* (193)
Gimpel, I., 84, *129* (226)
Görlich, P., *127* (194), *128* (219)
Gooden, F. S., 313 (31, 32), *316* (31, 32)
Goward, F. K., *316* (19)
Graham, R. L., 239, 265 (127), *267* (73), *268* (127)
Greenblatt, M. H., 81, *130* (256)
Greenstein, J. L., 379 (23, 26), *380* (23, 26)
Grenchik, R., 252, *267* (88)
Grosse, A. V., 258 (108), *268* (108)
Grove, D., 239 (68), *267* (68)
Gubanov, A., *126* (157)
Güntherschulze, A., 117, 118, 119, *121* (23, 24)
Gurewitsch, A. M., 296 (20), *316* (20)
Gurilev, B., *125* (130)
Gurney, R. W., 19 (38), 22 (44), *62* (18), *63* (38, 44), *64* (87)

H

Hachenberg, O., *129* (236)
Hagen, C., *122* (68), *124* (124)
Hagstrum, H. D., 265 (126), *268* (126)
Halpern, J., 309 (29), *316* (29)
Hammer, W., 250, *267* (79)
Hannay, N. B., 42, *64* (77)
Harang, L., *345*
Harries, J. H. O., *129* (234)
Hartman, C. D., 75, *127* (185)
Hass, W., 36, *63* (67)
Hastings, A. E., 89, *126* (158)
Haworth, L. J., 83, *121* (30, 46)

Hayden, R. J., 250, 252, 253, 258 (101–104, 106), 260 (110, 113), *267* (86), *268* (89, 101–104, 106, 110, 113)
Hayner, L. J., 80, 81, *121* (31), *122* (76)
Headrick, L. B., 98 (136), 111 (136), *125* (136)
Hecht, S., 148, *166* (10)
Hedvall, J. A., 25 (53), *63* (53)
Heiman, W., 98 (159), 113, *126* (159)
Heinze, W., 12, 36, *63* (28, 67, 70)
Hellivell, R. A., *345* (31)
Helmholtz, A. C., 258 (107), *268* (107)
Hemmendinger, A., 238 (57), *267* (57)
Henneberg, W., 197, 205 (19), *218* (11, 19)
Henyey, L. G., 379, *380* (23)
Herb, R. G., 271 (4), *315* (4)
Herbstreit, J. W., 364, *380* (27), 389, 419
Herold, K., *122* (69)
Herring, C., 7, 25, 62, *63* (57)
Herzog, R., 205 (18), *218* (18), 223, 225, 226, 246 (19), *266* (15, 19)
Hess, D. C., 250, 252, 253, 258 (102–104, 109), 260 (110, 113), *267* (86), *268* (89, 102–104, 109, 110, 113)
Hey, J. S., *345*, 356, 366, *380* (8, 11, 14)
Hickam, W., 239 (68), *267* (68)
Hickock, 114
Hide, G. S., 313 (31), *316* (31)
Higgins, G. C., 149, *166* (11)
Himpan, J., 170, *218* (1)
Hintersberger, H., 98 (125), 107, 109, 111 (125), *124* (125), 238 (53), *267* (53)
Hipple, J. A., 208, 211 (20), *218* (20), 222 (6), 239, 241, 247, 260 (6), 265 (124), *266* (6), *267* (63, 68, 72), *268* (124)
Hirano, K., 34, 35 (66), 41, *63* (66)
Honig, R. E., 235, *266* (43)
Hoover, H. H., 220 (5), 239, 260 (5), 262, *266* (5)
Huber, H., 5, 10 (13), 11, *62* (13)
Hudson, C. M., 271 (4), *315* (4)
Hulburt, E. O., *345*
Hustrulid, A., 220 (11), 239, 264 (11), *266* (11)
Hutter, R. G. E., 170, 183, 184 (7), 194 (7), 205 (17), *218* (3, 4, 7, 17), 226, *266* (23)
Huxford, W. S., 34, *63* (65)
Huxley, L. G. H., *345*
Hyatt, J. M., 94, *120* (11)

I

Iams, H. A., *121* (32), 153, 154, *166* (17, 18)
Ignatov, A. S., 108 (see Timofeev), *121* (40)
Inghram, M. G., 236, 239, 250, 251, 252, 253, 258 (101–105, 109), 260 (110, 113), 265 (123, 131), *267* (44, 70, 85, 86), *268* (89, 101–105, 109, 110, 113, 123, 131)
Ingram, L. J., *344* (8)

J

Jacobs, R., 264 (117), *268* (117)
Janes, *114*
Jansky, K. G., 348, 349 (3), 350, 364, *380* (1–3)
Jensen, H. H., 313 (32), *316* (32)
Joffe, J., 34, *63* (64)
Joffe, M. S., *127* (195)
Johannsen, G., 238, 246, *267* (55)
Johler, J. R., 364, *380* (27)
Johnson, J. B., 98 (257), 102, 103, 104, 105, 109, 112, *129* (235), *130* (248, 249, 257)
Johnson, R. P., 118, *122* (73)
Jones, L. A., 149, *166* (11)
Jones, T. J., 4, 5, *62* (10)
Jonker, J. L. H., 78, *123* (91–93), 124 (126)
Jordan, E. B., 224, 226, 238, 239, 243, 244, 245, 255, 256, *266* (16), *267* (65, 69), *268* (94)

K

Kadyschevitsch, A. E., 82, 83, 89, 92, 105, 110, *126* (160), *127* (196), *129* (242, 243)
Kamogawa, H., *126* (161)
Katz, H., 88, *122* (70)
Katzin, M., 419, *423* (28)
Kawamura, H., 31, 34, 35 (66), 40, 41, 51, *63* (61, 66)
Kennan, P. C., 379, *380* (23)

Kennard, E. H., 233 (42), *266* (42)
Kennedy, W. R., *129* (227)
Kerr, D. E., 397 (11), 402 (11), *422* (11)
Kerst, D. W., 278, 281, 286 (14), *315* (11, 13, 14)
Khan, A. W., 71, *127* (191)
Khlebnikov, N. S., 68 (95), 70, 79, 115, *123* (94, 95), *125* (127, 129)
Kingdon, K. H., 229 (33), *266* (33)
Kirvalidze, I. D., *126* (162)
Kluge, W., *122* (71)
Knoll, M., 74, *121* (47), *125* (128), *129* (236)
Knudsen, M., 233, *266* (39, 40)
Koch, J., 228, *266* (32)
Kohlman, T. P., 253, *268* (91)
Kollath, R., 66, 67 (72), 68 (72), 69, 74, 76, 78 (98), 83, 96, *122* (72), *123* (96–98), *126* (163), *127* (198)
Koller, L. R., 68 (250), 118, *122* (73), *130* (250)
Korshunova, A., 115, *123* (94), *125* (129)
Kosman, M., *125* (130)
Kramers, H. A., 379, *380* (25)
Krautz, E., 111 (123), *124* (123), *125* (131)
Krenzien, O., 105, *122* (74), *128* (220)
Kruithof, A. A., 114, *121* (34)
Kubetzkii, see Kubetzky
Kubetzky, L. A., *122* (75), *129* (237)
Kuljvarskaya, B. S., *130* (260)
Kundt, W., 68 (230), 75, 79, *125* (144, 145), *129* (229, 230)
Kurrelmeyer, B., 80, *122* (76)
Kushnir, Y. M., 82, *125* (132, 133), *127* (199)
Kwarzchawa, I. F., 115, *121* (48), *129* (238)

L

Lallemand, A., *130* (251)
Landon, D. H., 68 (108), *124* (108)
Lange, H., 94, *120* (6)
Langenwalter, H. W., 82 (33), 83, *121* (33)
Langmuir, I., 229 (33), *266* (33)
Langmuir, R. V., 296 (20), *316* (20)
Lantz, P. M., 258 (105), *268* (105)
Lapp, R. E., 260 (111), *268* (111)
Larson, C. C., 150, *166* (15)
Law, H. B., 155, *166* (19)
Lawrence, E. O., 271, 272 (7), 276 (10), 301 (18), *315* (6, 7, 10), *316* (18)
Leiger, E., 265 (119), *268* (119)
Levin, N. M., 119 (240), *129* (240)
Lincoln, J. V., 384, *421* (2)
Livingston, M. S., 272 (7, 8), 276 (8, 9), 277 (8), *315* (7–9)
Lofaren, E. J., 300 (26), *316* (26)
Loosjes, R., 44, *64* (80)
Lortie, M., *124* (112), *125* (150)
Lowry, G. F., 4, 58, 61 (7), *62* (7)
Lozier, W. W., 250, 251, *267* (83)
Lukjanov, S. J., 77, *123* (99), *129* (239)
Lunkova, J., 108 (177), *127* (177, 178)

M

MacColl, L. A., 84 (35), *121* (35)
McCready, L. L., 376, 377 (21), 379 (22), *380* (21, 22)
McIntosh, L. R., 271 (3), *315* (3)
McKay, K. G., 71, 72, 73, 79 (221), *128* (221)
MacKenzie, K. R., 300 (26), 301 (18), *316* (18, 26)
McMillan, E. M., 265 (132), *268* (132), 276 (10), 289, 295, 300, 301 (18), *315* (10), *316* (16, 18)
McNish, A. G., *344* (9), 384, *421* (2)
Mahl, H., 68 (134), 118, *122* (77), *123* (100) *125* (134)
Majewski, W., *125* (135)
Malter, L., 114, 116, 117, 118, 119, *121* (49), *122* (61), *128* (200)
Mann, A. K., 256, *268* (97)
Manning, L. A., *345*
Mariner, T., 220 (9), 260 (9), 265 (9), *266* (9)
Martin, S. T., 98 (136), 111 (136), *125* (136)
Martyn, D. F., *345*, 376 (17), *380* (17)
Massey, H. S. W., 6 (17), *62* (17), *344* (10)
Mathes, I., 116, *128* (201)
Matsumoto, T., *125* (143)
Mattauch, J., 225, 226, 238, 245, 246 (19), 247, 256, *266* (19), *267* (52, 53, 61)

Maurer, G., 107, 110, *128* (202)
Maurer, R. J., 6 (15), 13 (15), *62* (15)
Mecklenburg, W., 12, *63* (29)
Mendenhall, H. E., 68 (258), *130* (258)
Metcalf, G. F., 250, *267* (82)
Meyer, W., 14, *63* (32)
Meyerhof, W. E., 36 (71), *64* (71)
Miller, P. A., Jr., 81, *130* (256)
Miller, P. H., 36 (71), *64* (71)
Milyutin, I., *125* (132, 133)
Mimno, H. R., *344* (2), *345* (38)
Mitchell, J., 239, *267* (67)
Mitra, S. K., *344* (1)
Mohler, F. L., *345*
Moon, P. B., 228, *266* (27)
Moore, G. E., 58, 59, 62, *64* (88)
Morgulis, N. D., 39, *64* (74), 101, 102, 104, 107, *123* (101), *125* (137), *126* (164–166), *128* (204)
Morozov, P. M., 75, *128* (205, 206)
Morrison, J., 21, 32, *63* (41)
Morton, G. A., 112, 114, *122* (61), *126* (168), 150, 151, 153, 164, *166* (12, 16, 17)
Moss, F. K., 146, *166* (5)
Moss, H., 195, *218* (8)
Mott, N. F., 19 (38), 22 (44), 34, 47, *62* (18), *63* (38, 44, 63), *64* (87)
Moxon, L. A., 354, *380* (10)
Mühlenpfort, J., 119, *123* (102)
Mueller, C. W., 98 (244), 99 (244), 100, 111, *129* (244)
Müller, H. O., 76, 77, 106, *122* (78)
Mulliken, J., 237, *267* (45)
Murgoci, R., 4, 58, 61 (8), *62* (8)
Murphy, B. E., 265 (125), *268* (125)
Murphy, R. F., 265 (129), *268* (129)
Muskat, M., 210, *218* (26)
Mutter, W. E., 43, 44, 48, 58, 62, *64* (79, 81)
Myers, D. M., 95, *122* (79)

N

Nagorsky, A., 101, 102, 104, *123* (101)
Naismith, R., *344* (5, 7, 8)
Nechaev, I. V., *127* (195)
Nelson, H., 98 (103), 99 (103), 107, 111, 112, *123* (103, 104), *125* (138), *126* (167)
Nemilov, Y. A., 113, *128* (207)
Ney, E. P., 236, 250, 251, 252, 256, 265 (128) *267* (44, 85), *268* (97, 128)
Nichols, M. H., 73, *121* (37)
Nier, A. O., 220 (10, 11, 12), 224, 231, 235, 236, 238, 239, 241, 250, 251, 252, 254, 258 (108), 264 (11), 265 (10, 121, 125), *266* (10, 11, 12, 17, 37), *267* (44, 62, 70, 85), *268* (108, 121, 125)
Nijboer, B. R. A., 17, *63* (36)
Nishibori, E., 31, 40, 41, *63* (61, 66), *128* (203)
Norris, L. D., 265 (131), *268* (131)
Norton, K. A., 369, 376 (18), *380* (18), 390 (8), 392 (9), *398*, 405 (13, 14), 406 (19, 20), 407 (8, 13), 419 (23), 420 (23), *422* (8, 9, 13, 14, 19, 20), *423* (23)
Nottingham, W. B., 98 (80), 112, *123* (80)
Nyquist, H., 359, *380* (13)

O

OBryan, 9
O'Daniel, H., 25 (56), *63* (56)
Oliphant, M. L., 199 (15), 200, *218* (15), 228, 238, 247, *266* (27), *267* (54), 313 (31), *316* (31)
Omberg, A. C., 376 (18), *380* (18)
Overbeek, A. J. V., *123* (93)

P

Paetow, H., 120, *125* (139), *128* (208)
Palewsky, H., 252, *267* (88)
Palluel, P., *130* (258a–c)
Parker, G. W., 258 (104, 105), *268* (104, 105)
Parsons, S. J., 354, 356, 366, *380* (8, 11, 14)
Pawsey, J. L., 356, 376, 377 (21), 379 (22), *380* (16, 21, 22)
Payne-Scott, R., 376, 377 (21), 379 (22), *380* (21, 22)
Pekeris, C. L., *345*, 397 (12), 402 (12), *422* (12)
Penning, F. M., 114, *121* (34), 232 (38), *266* (38)
Pes'yatski, I. F., *125* (140)

Petry, R. L., 68 (7), *120* (7)
Phillips, J. W., 354, 356, 366, *380* (8, 11, 14)
Phillips, M. L., *345*, 386 (6), 420 (26, 27), *422* (6), *423* (26, 27)
Phips, T. E., 40 (75, 76), *64* (75, 76)
Picht, J., 170, *218* (1)
Pierce, J. A., *345*
Pierce, J. R., 80, *124* (107)
Pike, E. W., 116, *128* (215)
Pineo, V. C., *344* (9)
Piore, E. R., 112, 118, *123* (81), *126* (168)
Pollack, H. C., 296 (20), *316* (20)
Pollard, E., 220 (3), *266* (3)
Pomerantz, M. A., 98 (252), 102, 103, 104, 112, *130* (252, 253)
Popper, K., 238, 246, *267* (55)
Prescott, C. H., 21, 32, *63* (41)
Pyatnitski, A. I., 108, 115, *121* (53), *123* (82), *124* (105a), *127* (178)

R

Rajchman, J. A., *126* (154)
Rakov, V. I., *125* (141)
Rall, W., 239, *267* (76)
Ramberg, E. G., 150
Ramsey, 52
Randenbusch, H., 117, *128* (209)
Randmer, J., *129* (236)
Rann, W. H., *125* (142)
Rao, S. R., 74, *120* (17)
Rapuano, R. A., 309 (29), *316* (29)
Rayleigh, J. W. S. (Lord), 406 (18), *422* (18)
Reber, G., 349, 350, 354 (7), 356, 376, 379 (26), *380* (4, 5, 7, 26)
Reeves, P., 146, *166* (8)
Reichelt, W., 76, *126* (169)
Reiman, S. P., 220 (10), 265 (10), *266* (10)
Reimann, A. L., 4, 58, 61 (8), *62* (8)
Richardson, J. R., 300 (26), *316* (26)
Richardson, O. W., 72, 84, *120* (18), *129* (226)
Rittenberg, D., 265 (118), *268* (118)
Rogers, F. T., 222, *266* (14)
Rooksby, H. P., 9, 10 (21), 24 (48–50), *62* (20, 21), *63* (48–50)
Rose, A., 39, *64* (73), 144, 148, 154, 155, 160, *165* (1), *166* (9, 18–20)
Rothe, H., 29, *63* (59)
Rudberg, E., 81 (50), 82 (50), 83, 84, 85, 89, 91, 105, *121* (50, 51)
Rüdenberg, H. G., 183, *218* (6)
Ruedy, J. E., 116, *128* (215)
Rumbaugh, L. H., 238, *267* (56, 58)
Rustad, 251
Rydbeck, O. E. H., *345*

S

Saegusa, H., *125* (143)
Salisbury, W. W., 276 (10), *315* (10)
Salow, H., 90 (171), 98 (170), 106, 112, *126* (170, 171)
Salzberg, B., *121* (32)
Sampson, M. B., 238, *267* (59)
Sander, K. E., 354, *380* (9)
Sandhagen, M., *123* (105)
Sard, R. D., 80, *130* (254)
Schade, O. H., 144, 153, 165, *166* (4)
Scherer, K., 98 (172), 106, 112, *126* (172)
Schlechtweg, H., *129* (228)
Schlesinger, K., 196, 197, *218* (9, 10)
Schmerwitz, G., 79 (14), *120* (14)
Schmidt, A., 14, *63* (32)
Schneider, E. G., *124* (106)
Schnitger, H., 117, *126* (173)
Schottky, W., 29, 34, 47, *63* (59, 60)
Schroeder, A. C., 141, *165* (1)
Sears, R. W., 4, *62* (9)
Seaton, S. L., *344* (12, 14)
Seifert, R. E., 40 (75), *64* (75)
Seitz, F., 6 (14), 13 (31), 20 (39), 22 (45), *62* (14), *63* (31, 39, 45), *64* (83)
Serber, R., 286 (14), 301 (18), *315* (14), *316* (18)
Sewell, D. C., 301 (18), *316* (18)
Sharpe, J., 199 (13), *218* (13)
Shaw, A. E., 230, 239, 258 (109), *266* (34), *267* (76), *268* (109)
Shire, E. S., 199 (15), 200, *218* (15), 238, 247, *267* (54)
Shockley, W., 80, 110 (36), *121* (36), *124* (107)
Shull, F. B., 208, 217, *218* (25)
Shultz, E. L., *345*
Siegbahn, K., 208, *218* (22, 24)
Simpson, K. M., 301 (18), *316* (18)
Siri, W., 239, 252, *267* (77)

Sixtus, K., 71, 72, *120* (15)
Skellett, A. M., *128* (222)
Skinner, 9
Slater, J. C., 84, 91, *121* (51), 306, 309 (28, 29), 310, *316* (28, 29), 370 (28, 29), *380* (28, 29)
Slattery, J. J., *346* (44)
Sloan, D. H., 306, *316* (27)
Slobrod, R. L., 256, *268* (96)
Smith, D. T., 220 (4), 260 (4), *266* (4)
Smith, L. G., 250, 251, *267* (83)
Smith, L. P., 228, *266* (31)
Smith, N., *346*
Smith, P. T., 231, 238, 250, 251, *266* (36), *267* (83)
Smoluchowski, M., 233, *266* (41)
Smyth, J. B., 405, 407 (16), 419 (16), *422* (16)
Smythe, W. R., 231, 238, 247, *267* (51, 56, 57, 64)
Soller, T., 83, *120* (19)
Sorg, H. E., *130* (245)
Southworth, G. C., 376 (19), *380* (19)
Sperduto, A., 271 (3), *315* (3)
Sproull, R. L., 32, 38 (62), 39, 51, 52, 55, *63* (62)
Stehberger, K. H., 86, *120* (12)
Stephens, W. E., 226, 239, *266* (24), *267* (74)
Stevens, C. M., 220 (11), 239, 264 (11), *266* (11), 267 (70)
Stewart, G. S., *345* (28)
Stewart, J., 384, *422* (5)
Stewart, J. Q., *346*
Steyskal, H., *122* (71)
Straker, T. W., *346* (49)
Straus, H. A., 239, 251, *267* (66)
Strübig, H., *121* (52)
Suhrmann, R., 68 (230), 75, 79, *125* (144, 145), *129* (229, 230)
Swank, R. K., 252, *267* (88)
Swartholm, N., 208, 218 (22–24)
Symonds, J. L., 313 (32), *316* (32)
Sziklai, G. C., 141, *165* (1)

T

Tanaka, M., *128* (210)
Tate, J. T., 220 (4), 231, 238, 260 (4), *266* (4, 36)
Taylor, J. E., 239, *267* (71)
Teichmann, H., *128* (211)
Tellegren, B. D. H., *120* (8)
Terrill, H. M., 88, *120* (2)
Teves, M. C., *123* (91), *126* (174)
Theile, R., 74, *125* (128)
Thode, H. G., 239, 265 (127), *267* (73), *268* (127)
Thomas, H. A., 239, *267* (72)
Thomson, B. J., 250, *267* (82)
Thomson, J. J., 220, 228, 232, 237, 238, 250, *265* (1), *266* (25), *267* (78, 80, 81)
Thornton, R. L., 276 (10), 301 (18), *315* (10), *316* (18)
Tilton, E. P., 387 (7), 422 (7)
Timofeev, P. V., 73, 108, 115, 116, *121* (39, 40, 53), *122* (62), *123* (82), *125* (146), *126* (175, 176), *127* (177–180), *130* (246)
Tol, T., 88, *121* (38),
Townes, C. H., 379, *380* (24)
Treolar, L. R. G., 68 (108), 72, 73, 74, 92, 93, 95, 115, *122* (54), *123* (83), *124* (108, 109)
Trey, F., 107, *129* (231)
Trolese, L. G., 405, 407 (16), 419 (16), *422* (16)
Truell, R., 90, *128* (223)
Trump, J. G., 71, 86 (259), 87, *130* (259)
Tscheischwili, L., 25 (56), *63* (56)
Turnbull, D., 40 (76), *64* (76)
Turnbull, J. C., 85 (110), *124* (110)
Turner, C. M., 271 (4), *315* (4)
Tyagunov, G. A., *126* (153)

U

Unwin, J. J., *344* (10)
Urey, H. C., 265 (119), *268* (119)

V

Van de Graaff, R. J., 71, 86 (259), 87, *130* (259), 271 (2, 3,), *315* (2, 3)
Van Horn, J. R., 260 (111), *268* (111)
Varadachari, P. S., *127* (182)
Vaughan, A. L., 220 (4), 260 (4), *266* (4)
Veenemans, C. F., *9*, *62* (22)
Veksler, V., 289, 295, *316* (17)

Venkatamaran, K., *344* (15)
Verway, E. J. W., 6 (16), 25 (16), *62* (16)
Vick, F. A., 2, *62* (4)
Villard, O. G., *345* (31)
Vink, H. J., 44, *64* (80)
Vudynski, M., 98 (145), 100, 105, 106, *124* (111), *125* (147–149), *128* (212)

W

Wagener, S., 5, 10 (13), 11, 12, 36, *62* (13), *63* (28, 70)
Waldmeier, M., 384, *422* (4)
Walkinshaw, W., 397 (10), 402 (10), *422* (10)
Wall, R. F., 228, 247 (30), *266* (30)
Walton, E. T. S., 271 (1), *315* (1)
Wang, C. C., 81, *130* (247)
Warnecke, R., 68 (56), 71, 72, 79, 93, *122* (56, 57), *123* (84), *124* (112), *125* (150)
Warren, R. E., 271 (4), *315* (4)
Was, D. A., 88, *121* (38)
Washburn, H. W., 220 (5, 7, 8), 239, 260 (5, 7, 8), 262, *266* (5, 7, 8)
Wecker, F., *128* (213)
Weimer, P. K., 155, *166* (19)
Weiss, G., 114, *122* (58)
Weiss, J. G., 116, *127* (192)
Wells, H. W., *344* (12), *346*
Wendt, G., 170, 200, 212, *218* (2, 16)
West, S. S., 238, *267* (56)
Westendorp, W. F., 281 (12), *315* (12)
Whalley, W. B., 239, 250, 252, *267* (75)
Whiddington, R., 88, 90, *120* (1)
White, F. W. G., *346*
Widell, G. G., 9, *62* (23)
Wideroe, R., 271 (5), 306 (5), *315* (5)
Wiley, H. F., 220 (7), 260, (7), *266* (7)
Williams, D., 265 (130), *268* (130)
Williams, T. W., 239, *267* (72)
Wilson, D. W., 220 (10), 265 (10), *266* (10)
Wilson, H. A., 2, 6, *62* (2)
Wise, E. M., 17 (35), *63* (35)
Wolff, H., *128* (214)
Wooldridge, D. E., 68 (152), 71, 73, 74, 75, 76, 91, 92, *125* (151, 152), *127* (184, 185)
Wooten, L. A., 21, 32, *63* (42, 43)
Wright, B. T., 300 (26), *316* (26)
Wright, D. A., 24 (51), 26 (51), 34, 43, 47, 52, *62* (5), *63* (51)

Y

Yasnopol'ski, N., 107, *126* (153), *127* (186)
Young, W. S., 262, *268* (115)
Yumatov, K. A., 108 (180, 181), *127* (180, 181)
Yuster, P., 265 (130), *268* (130)

Z

Zahl, H. A., *346* (44)
Zernov, D. V., 119, *129* (240, 241), *130* (260)
Ziegler, J. A., 239, 265 (127), *267* (73), *268* (127)
Ziegler, M., 80, *122* (59, 60)
Zuhr, R., 264 (117), *268* (117)
Zworykin, V. K., 114, 115, 116, *122* (61), *126* (154), *128* (215), 150, 151, 153, *166* (12, 13, 16, 17)

Subject Index

A

- Aberration,
  - lens, 200
  - second order – of mass spectrometer, 207
- Absorbing area, 371, 372
- Absorption,
  - atoms, 5
  - characteristic, 22
  - ground –, 357
  - ionospheric, 351, 356, 366, 368
  - light in retina, 148
  - light quanta, 136, 137
  - mesons, 269, 270
  - neutron – cross section, 260
  - radio waves, 333
  - secondaries, 71, 76, 77
    - before emission, 91
  - spectra, 22
- Accelerating fields, emission in, 37 ff.
- Accelerator,
  - induction, 278 ff.
  - ion, 270
  - linear, 306 ff.
  - magnetic, 270, 271
  - particle, 269 ff.
  - resonance, 271 ff.
- Activation,
  - of cathode, 3
  - density of – centers, 14
  - energy,
    - optical, 22
    - thermal, 13
  - state of, 17
- Activators, 23
- Active centers, 108
- Adatom, 5
- Adsorption,
  - of monolayer, 79
  - Van der Waals', 75
- Affinity, electron –, 8, 12, 41
- Age of earth, 265
- Age of elements, 265
- Air refraction, 395
- Alkali halides, 98
- Alkaline earth,
  - carbonates of the – elements, 3
  - compounds, 98
  - oxides, 2, 9, 23, 58
- Aluminum oxide, 98, 107, 117, 118
- AM channels, 382
- Amplifier,
  - beam deflection –, 167
  - electronics, 251
  - low noise figure – circuit, 363
  - noise of, 133
  - vibrating reed type of, 252
- Amplitude modulation (AM), 381
- Analyzer, mass, 220, 226, 237
  - magnetic, 81, 96, 222
  - symmetrical, 223, 226
- Angular, error, 446
- Angular size of test object, 139
- Anode, hot – source, 228
- Antenna,
  - characteristics, 429
  - directivity pattern of, 357
  - gain of directional, 373
  - half-wave dipole, 349, 357
  - loop –, 427 ff.
  - noise of, 366
  - Q of, 440, 442
  - radar, 362
  - radiation, 369, 370
  - receiving –, height of, 390
    - directional –, 348
  - rhombic, 348
  - simple threetower –, 431
  - transmitting –, 429
    - FM –, 391
    - height of –, 390
- Aperture, effect of small –, 311
- Atmospheric,
  - ducts, 397, 398
  - noise, 339, 340, 434
- Atoms, absorption of, 5
- Attenuation,
  - field intensity, 418
  - radio wave –, 397
- Auroral displays, 332
- Automatic flight, 435
- Available power in radio reception, 358
- Avalanche of electrons, 106
- Azimuth,
  - determination of, 430

  finder, 429
  finding, 427, 431

B

Back diffusion, 233
Background counts of amplifier, 133
Bands,
  conduction –, 7, 22, 35
  density of electrons on conduction –, 13
  electron –, 92
  energy –, 6
  nature of filled –, 9
Bandwidth, 163, 164
  effective noise –, 359
  radio signal, 437
Barium, 4, 5, 25
  BaO, 6 ff.
  (BaSr)O, 6 ff.
  $BaCO_3$, 3, 9, 25
  Ba_2SiO_4, 24, 25, 27, 28
  excess – content, 21
  radioactive –, use of, 22
Barrier layer, 33
  penetration, 47
  theory, 25
Baseline, 446, 449
Base metal, 61
Beam focusing effect, 168
Beam-splitter, 304
Beryllium target, 74
Beta-ray spectrometer, 168
Betatron, 278 ff.
  accelerating cycle, 279, 280, 281
  critical frequency, 288
  orbit radius, 284
  oscillation, 286, 293, 294
  path of electron, 278
  radiation loss, 295
  rotation frequency, 284
  start, 296
  transformer principle, 287
Biological tracers problems, 265
Black body radiation, 369 ff.
Black surface of secondary electron emitters, 77
Bragg law, reflection, 85
Break down, 3
Brightness of scene (television), 135 ff.
Broadcast band, FM, 381 ff.
  propagation in, 381 ff.
Burst duration (shot effect), 80

C

Calcium oxide, 9, 23, 98, 105
Cathode,
  activation of, 3
  complete –, 29
  electrolytic base –, 43
  emission theory, 4 ff.
  flushing, 55
  life, 12
  oxide coated –, 1 ff.
    examination of surface of emitting –, 12
    preparation of, 3
    properties of, 3
  sparking, 55
  surface of, 5
  thermionic emission of, 2, 29 ff.
Cathode ray,
  electrostatic – tubes, 195
  indicator, 430
  screen, 23
  tubes, 167, 196
Charged particles, deflection of, 167 ff.
Charge pattern, 157
Chemical,
  barrier, 34
  interface barrier layer, 43
  potential, 13
    of electron, 7
    of semiconductors, 29
  tracers problems, 265
Circular system, 436, 444
Coated cathodes, oxide (see under "Cathode")
Coating,
  conductivity, 17, 43
  evaporation of, 59
  fluorescence of oxide cathode –, 23
  heat treatment, 27
  monolayer, 59
  properties, 2, 3
Co-channel FM signal, 412
Collector,
  multiple –, 251
  width of – slit, 248
Collimated beam, 223
Composite surfaces (secondary electron emission), 114 ff.
Compton electrometer, 251
Conductivity, electronic, 12 ff., 27, 40 ff.
  atmosphere, 317
  coating –, 17, 43
  discontinuities in, 431
  earth, 436
  after exposure to light, 19
  interface –, 15, 43
    activated –, 27, 28
  material, 7
  measurement of, 14

- temperature dependency, 13, 15, 16, 40
- after thermal activation, 14

Conduction band, 7, 22, 35
- contribution of electrons to, 9
- density of electrons on, 13

Cones of the retina, 147
Contamination of surface, 67
Continuous wave,
- system, 434
- technique, 318

Contrast, 136, 139
Converter, low noise figure, 363
Cosmic radio noise, 347 ff., 388
- frequency law, 354, 357, 374, 375
- intensity, 354, 357
- measurement, 354
- origin of, 379

Counting of light quanta, 132
Critical frequency, 318, 330
Cross section, neutron absorption, 260
Cross-talk, 172
Curvature,
- earth, 394
- radio path, 402
- transmission path, 392

Cyclotron, 168, 271 ff.
- Berkeley 184-inch, 302, 303
- electric focusing, 273, 276
- dimensions, 301
- frequency modulated, 289
- limiting energy, 276
- magnetic focusing, 275
- path of ions, 272

Cygnus, constellation, 349
- point source in, 356

D

Dead reckoning, 426
Decca system, 434, 436
- modulated –, 450

Deflection,
- beam of charged particles, 167 ff.
- defect, 172
- electrostatic – fields, 168 ff.
  - crossed superimposed, 176
  - crossed unbalanced, 188
  - single balanced, 174
  - single three-dimensional, 185
  - two crossed balanced, 173
  - two crossed unbalanced, 174
- electrostatic – system, 196, 215
- fields, 168 ff.
  - improved, 195
  - two-dimensional, 173
- large-angle –, 200 ff.
- magnetic – fields, 190
  - crossed, 190
  - single, 175
  - two crossed, 174
- small-angle –, 168 ff.

Defocusing effect, 188, 196
Detailed balancing, principle of, 370
Detector, 220
- leak –, 264

Diffraction, X-ray, 5, 9, 11, 24, 26
Diffusing mixing term, 233
Diffusion, back –, 233
Dilution factor, 445
Diode noise source, 360
Dipole field, magnetic, 197
Direction,
- error, 428
- finder,
  - aircraft automatic –, 429
  - crossed loop –, 429
  - instantaneous, 430
  - spaced loop –, 430
- finding, 427 ff.
  - radial system for, 436
  - system, 446
  - on transmitting station, 431
- focusing (see under "Focusing"), 223 ff.

Directional,
- devices, 427
- focusing in mass spectrograph, 200
- gain of – antenna, 373
- receiving antenna, 348, 390

Discharge,
- gaseous – in magnetic field, 232
- gaseous – type of source, 228
- lightning, 348
- tube electron source, 276

Discrete energy states (levels), 6
Distance-measuring,
- equipment, 448
- system, 444, 445

Distortion, 189
- – pattern, 183
- pattern –, 193, 194
  - correction of, 195 ff.
  - reduction of, 197
  - trapezoidal, 195
- received field, 428
- spot –, 183 ff.
  - correction of, 195 ff.
  - reduction of, 196

Distribution of field intensity, 401
Disturbances, forecasting of, 330
Doppler shift, 341

Double focusing (see under "Focusing"), 225
Double layer formation, 106
Doublet, 257
D region, 324
Dynamic,
  methods for secondary emission, 111 ff.
  two gun method, 113
Dynatron characteristic, 101

E

Eastman III-O plates, 250
Eclipse, artificial, 331
Effective,
  power of radio station, 413, 415, 417
  temperature, 366, 369, 370
E layer, 324
  sporadic – interference, 420
  transmission, 443, 446
Electrode-to-screen distance, 196
Electrolytic base cathode, 43
Electrometer, 251
Electron,
  affinity, 8, 12, 41
  avalanche, 106
  bands, 92
  beam, 95, 134
    point-focused –, 196
    voltmeter, 112
  bombardment source, 231
  chemical potential of, 7
  conduction –, 13
  diffraction technique, 5, 11
  discharge tube – source, 276
  energy of, 7
  gun, 95, 196
  image, 153
  mean free path, 13, 20, 93
  mean mobility, 13
  multiplier, 251
    gain of, 142, 151
    multi-stage –, 161
    noise current, 80
  -optical picture, 74, 118
  path, 177, 188, 205, 278
  primary – (see under "Primary"), 66 ff.
  prism, 167
  projection tube, 12
  scattering, 92
  secondary – (see under "Secondary"), 65 ff., 152
  transmission of thin foils, 88
  trapping of, 8, 18, 19, 24
Electronic,
  amplifier, 251
  conductivity (see under "Conductivity")
  navigation system, 425, 427
Electrostatic,
  cathode ray tubes, 195
  deflection, 215
  deflection fields (see under "Deflection, electrostatic"), 168 ff.
  deflection system, 196, 215
  field,
    combination of magnetic and –, 225
    retarding –, 83
    uniform –, 168
  generator, 271
  lens, 224
Elektra system, 431
Emission,
  in accelerating fields, 37 ff.
  cathode – theory, 4 ff.
  decay phenomena, 51
  phosphorescence, 23
  photoelectric, 34
  primary, 58
  process, 42
  pulsed –, 32
    enhanced –, 32
    method, 51
  in retarding fields, 35 ff.
  Schottky –, 47, 49
  seat of, 4, 5
  secondary – (see under "Secondary electrons"), 58, 65 ff.
  secondary – coefficient, 58, 71
  thermionic, 2, 29 ff.
    of cathodes, 2 ff.
    direct current –, 9, 21
Emulsions, photographic, 250
Energy,
  bands, 6
  collector, 349
  discrete – states, 6
  of electron, 7
  kinetic, 274
  loss, 92
Equation,
  motion of electron, 175 ff.
  relativistic – of motion, 282
Equilibrium orbit, 280
Equisignal zone (lane), 431
E region of ionosphere, 387
Error,
  angular, 446
  orientation, 429
  quadrantal, 430

Euler-Lagrange differential equation, 176, 204
Evaporation rate, 237
Excitation,
  by infra red, 111
  optical, 22
Exposure time, 136, 139, 143
Extraordinary ray, 325
Eye, electric, 132
Eye (human), 132, 145 ff.
  angular size, 146
  exposure time, 143
  performance, 145 ff.
  sensitivity, 160
  storage time, 146
  threshold contrast, 146
  threshold signal-to-noise ratio, 146
  variable gain element, 148

F

Fading, 395, 409
Faraday cage, 96
Fathometer, 426
F center, 8
Fermi distribution function (level), 8, 29
Field,
  distribution, 170
  emission, 66
  enhanced emission, 99, 100, 106
  fringe –, 169
  thin film – emission, 100
Film,
  graininess of, 159
  noise of, 159
  photographic (see under "Photographic film"), 148 ff.
  thin oxide – phenomena, 58 ff.
Filters, velocity, 242
Finding, direction, 427 ff.
Fission product, 259
Fix, error of, 441 ff.
Fixing, 426
F layer of ionosphere, 324, 384
  interference, 387, 420
  transmission, 443
F1 layer of ionosphere, 320, 322
F2 layer of ionosphere, 320 ff.
Fluctuation,
  human eye, 148
  noise, 146
  sky-waves, 449
Fluorescence of oxide cathode coating, 23
Flux bars, 296
FM (frequency modulation),
  broadcast band, 381 ff.
  broadcast reception, 369, 390
  co-channel – signal, 412
  interference range of – broadcast, 410
  optimum frequency for – broadcasting, 414
  optimum separation of – stations, 413, 414
  propagation in the – broadcast band, 381 ff.
  transmitting antenna, 391
Focal length, determination, 213
Focusing,
  action on electron and ion beam, 199, 200
  beam – effect, 168
  direction –, 223 ff.
    electrostatic analyzer, 224
    electrostatic field, 225
    magnetic analyzer, 237, 240, 246
    magnetic field, 225
    in radial electrostatic field, 224
  double –, 225
    analyzers, 243
  electron gun, 196
  magnetic field, 280
  in mass spectrometer (see also under "Spectrometer, focusing"), 208
  phase –, 307, 308
  properties, 223
  two-directional, 212
  velocity –, 222, 240
    spectrograph of Aston, 240
Forces, nuclear, 269, 270
Forecasting of disturbances, 330
Free radical phenomena, 265
Frequency,
  critical, 318, 330
  high – transmission conditions, 384
  law of cosmic radio noise, 354, 357, 374, 375
  maximum –, 384
  modulated cyclotron, 289
  modulation broadcasting (see under "FM")
  of radio signal, 437
  transmission –, 390
Fringing field, 169, 183, 200, 222, 304

G

Gain of directional antenna, 373
Galactic radio noise, 348
Galaxy, 348
Gas analysis, 260
Gaussian path, 205
Gee system, 432, 433

Generators, direct voltage, 271
Geomagnetic,
  effect, 321, 322
  field, 322
  latitude, 322
  storms, 330
Geometrical error, 445, 446
Glasses, secondary emission of, 98
Graininess,
  of film, 159
  of motion picture, 161
Grain size of photographic film, 149
Grid effect, 99, 100
Grid method, 112
Ground, 98
  absorption, 357
  wave,
    field, 391
    field intensity, 418
    pattern, 443
    propagation, 390
    system, 436
    transmission, 428, 442
      efficiency of, 434
      phase measurement of, 442, 450
Gyromagnetic frequency, 325

H

Hall coefficient, 19, 20
Hamilton's principle, 203
Heat,
  conduction method, 111
  treatment, 27, 68, 69, 78
Helium, 264
High frequency transmission conditions, 384
High voltage particle acccelerator, 269
High voltage X-ray machines, 270
Holes, 8
Hydrogen evolution method, 21
Hyperbolic,
  grid-laying device, 432
  problem, 445
  stations, 446
  system, 434 ff.
    for time-difference method, 436

I

Iconoscope, 151 ff.
  image –, 153
  performance curve, 153
Ideal performance of television pickup tube, 135 ff.
Ignition system of automobile, 388
Image,
  dissector, 150 ff.
  iconoscope, 153
    performance curve of, 153
  orthicon, 155
Impurity, 10
  center, 23
  concentration, 18
  content, 3
  density of, 13, 21
  excess – semiconductor, 6
  levels, 18, 19, 21
  semiconductor of N type, 6, 7, 9
  type of semiconductors, 6
  vacant – levels, 17
Induction accelerators, 278 ff.
Infra red, excitation by, 111
Inhomogenities of surface potential distribution, 99
Insulators, 7, 97, 105 ff.
Interface,
  chemical – barrier layer, 43
  conductivity, 15, 43
    activated –, 27, 28
  contact, 44
  layer, 24, 32
    color of, 26
    measurement of thickness of, 26
  properties of, 2, 24 ff.
  silicate –, 26
  thickness, 25
Interference,
  F layer –, 387
  -free service areas, 415
  between radio stations, 382
  range of FM broadcast, 410
Intrinsic semiconductor, 7
Ion,
  accelerators, 270
  beam source, 220
  density, 318, 320, 321
  detection, 248, 254
  detector, 250, 252
  Phillips – gauge, 232
  positive, 95
  prism, 167
  sources, 228
  trajectories in magnetic and electric fields, 221
  trap, 167, 199
Ionic decay, 324
Ionization,
  atmosphere, 437
  degree of, 229
  efficiency, 229
  fractionation in – process, 254

  solar – radiation, 328
Ionizing, solar – agent, 322
Ionosphere, 318, 382
  action, 428
  E region, 387
  F region, 324, 384
  maximum ion density in, 318
  reflection from, 384
  sudden – disturbances (SID's), 364
Ionospheric,
  absorption, 351, 356, 366, 368
  forecast of – propagation, 384
  long distance – propagation, 382
  measurements, 387
  prediction of disturbances, 330
  refraction, 356
  research, 317 ff.
  sporadic – transmission, 387
  storms, 330
  waves, 405
Iso-butane, 260
Isotopes, 249
  determination of mass of radioactive –, 258
  exact masses, 256
  existence, 253
  neon, 253
Isotopic abundances, 253

K

Kinescope, 141, 145
  visual impression, 144
Kinetic energy, 274
Klystron, reflex, 81

L

Lambert distribution, 137
Langmuir-Child space charge,
  emission, 56
  relationship, 37
LaPlace's equation, 170, 202
Lattice,
  defect, 8
  vibration, 22
Layer,
  formation, 234
  thickness, 88
  E — transmission, 443, 446
  F — transmission, 443
  sloping –, 428
Leak,
  detection, 264
  mass flow – system, 236
  molecular flow –, 234
Lens,
  aberration, 200
  diameter, 139
  electrostatic, 224
  parameter, 136
  rotational symmetrical, 200, 214
  television pickup tube, 137
  two-dimensional diverging, 197
Light quanta, 132, 133, 147
  absorption, 136, 137, 148
  counting, 132
Light spot scanner, 141, 142, 145
Linear accelerator, 306 ff.
  operation mode, 311
  phase stability, 307
Line width, 226
Loop antenna, 427 ff.
Loran system, 433, 442
  low frequency –, 422, 433, 449
  sky waves, 449
  transmission range, 433
Low current method, 111
Lowering of potential barrier, 5
Luminescence in semiconductors, 23

M

Magnetic,
  accelerators, 270
    resonance –, 271
  analyzer, 81, 96, 222
    symmetrical –, 223, 226
  deflection, 216, 221, 222
  deflection field (see under "Deflection, magnetic"), 190
  dipole field, 197
  field,
    combination of electrostatic and, 225
    control of, 252
    longitudinal, 83, 96
    modulation, 291
    transverse, 83, 96
    uniform, 168
  focusing, 154
Magnetron, 55
Malter effect, 66, 100, 106, 117
Mass,
  analyzer, 220, 226, 237
    magnetic, 81, 96, 222
    symmetrical, 223, 226
  discrimination effects, 232
  dispersion, 226
  flow leak system, 236
  flow principle, 235
  spectrograph (see under "Mass spectrometer")

  spectrographic measurement, 254, 255
  spectrometer, 168, 199 ff., 219 ff.
    aberration, second order, 207
    commercially available –, 265
    crossed field –, 205, 208
    focusing (see under "Focusing, spectrometer"), 208
    leak detection –, 264
    resolving power of, 241, 243
    use of, 253
  spectroscopy, 219 ff.
Maximum frequency of FM broadcasting, 384
Mean free path,
  electron, 13, 20, 93
  molecules, 233
Measuring angles, 427
Mercury, 260, 261
Mesons,
  absorption, 269, 270
  production, 312
Metal,
  base –, 61
  parabolic sheet – mirror, 356
  pure –, 67
  secondary emission yield, 68
Meteor,
  observation of whistle from, 341
  reflection from – trails, 340, 341
Micro-analyzer, 168
Microphotometer, 253
Microscope, scanning electron –, 168
Milky Way, 356
Modulation,
  amplitude – (AM), 381
    channels, 382
  frequency – (see under "FM")
  magnetic field –, 291
Molecular,
  doublet, 257
  flow of gases, 233
  flow leak, 234
Monolayer (Monoatomic layer), 78
  adatoms, 5
  adsorption of, 79
  coating –, 59
  theory of emission, 4 ff.
Mosaic of television pickup tubes, 133
Motion,
  equation of, of electron, 175 ff.
  graininess of – picture, 161
  of particles, 201
Multifrequency recorder, 342
Multiple charge doublet, 257
Multiple radio range, 431
Multiplier,
  electron –, 251
    noise current of, 80
  voltage – circuit, 271

N

Navigation,
  accuracy of – system, 443
  aids to, 427
  maximum accuracy of radio aids to, 443
  radio aids to, 427
    accuracy of, 448
  range of – system, 436, 437, 439
  standard – system, 435
Navigator, 425
  indicator oscilloscope of aircraft –, 436
Neodymium, 248, 249, 258
Neon, isotope, 253
Neutron absorption cross section, 260
Nickel, 3, 61, 71, 90
Night effect, 428
Noise,
  amplifier –, 133
  antenna, 366
  atmospheric, 339, 340
    level, 434
  cosmic radio – (see under "Cosmic")
  current, 152, 153
    electron multiplier, 80
  diode method, 359
  diode for very high frequency, 362
  external, 419
    factor, 366
    measurement of, 360
  extraterrestrial, 348
  figure, 358, 390
  incident – radiation,
    intensity, 371
    measured –, 372 ff.
  internal, 338
  low – figure,
    amplifier circuit, 363
    converter, 363
    Wallman –, 363
  natural, 338
    in function of wavelength, 437
  power, 371, 372
  -producing electrification of atmosphere, 437
  radio – (see under "Radio noise")
  random –, 347
  receiver –, 358, 419
    figure, 367
  reception of, 348
  solar, 339, 350

  suppressor, 388
  television circuit, 162
  temperature, effective, 361, 370
  thunderstorms, 340
  visibility, 160
  wide-band receiver, 440
Non-separative flow, 232
N type, impurity semiconductor, 6, 7, 9
Nuclear forces, 269, 270

O

Omnidirectional,
  beacon, 435
  range, 448
Omnirange, 435
Optical,
  activation energy, 22
  curved – axis, 210
  excitation, 22
  spectrograph, 263
Ordinary ray, 325
Orfordness beacon, 429, 430
Orthicon, 154, 158
Oscilloscope, indicator – of aircraft navigator, 436
Oxide,
  alkaline earth –, 2, 9, 23, 58
  cathode, 1 ff., 98
  coated cathode (see under "Cathode, oxide coated"), 1 ff.
  coating, 3
  thin – film, 2 ff.
    phenomena, 58 ff.
Oxidized targets, 116

P

Packing fraction, 232, 256
Parallel cylinders, 183, 184
Parallel plates with fringing field, 183
Particle accelerators, 269 ff.
Patch effect, 39
Pattern distortion, 193, 194
  correction of spot and, 195 ff.
  reduction, 197
  trapezoidal, 195
Performance,
  curves for pickup devices, 147 ff.
  ideal (see under "Ideal performance")
Phase,
  comparator, 430
  comparison, sensitivity of, 430
  focusing, 307, 308
  stable orbits, 289
  stability, 270, 289
  system, 440
Phillips ion gauge, 232
Phosphorescence,
  decay of – emission, 23
  semiconductors, 23
Phosphors, secondary emission yiel
Photocathodes, 114
  conducting –, 151, 153, 156
  multiplier –, 145
Photo cell, 141
Photoconductivity, 19
Photoelectric,
  emission, 34
  threshold, 35
  work function, 34
Photo electrons, 134
Photographic,
  emulsions, 250
  film, 148 ff.
    grain size, 149
    optimum performance, 149
    sensitivity, 160
  process, 165
Photoionization, 341
Photometric measurement, 263
Photomultiplier, 141, 145
Photosensitivity, 114, 115
Pickup tubes, television (see un "Television pickup tubes"), 131 f
Picture formation, television, 132, 133
Pilotage, 425, 426
  visual, 432
Platinum, 61, 89
Poiseuille's law, 233
Polarization,
  extraterrestrial radiation, 378
  radio signal, 428, 430
Polycrystalline surface, 70
Potential barrier, lowering of, 5
Power, effective, 413
Primaries (primary electrons), 66 ff.
  emission, 58
  energy, 67, 71
    loss of, 91
    rate of loss of, 87, 88
  range of – and secondaries, 87
  reflected, 71, 91
    directional scattering of, 75
    elastically –, 84
Principal planes, location of, 214
Prism,
  electron, 167
  ion, 167
Probe targets, 277
Propagation,
  in FM broadcast band, 381 ff.
  long distance ionospheric, 382

  radio – prediction, basic, 387
  velocity of – of radio waves, 450
    pe semiconductor, 6

    atching system, 442
  ystem, 439, 440
    in function of frequency, 439
  technique, 318
  transmission, error of, 442
Pulsed emission, 32, 51
  enhanced –, 32
  method, 51
upil diameter, 146

Q

Q of antenna, 440
  effect of, 442
Quanta, light (see under "Light quanta")
Quantum,
  theory, 6
  yield, 139
uarter-wave cavity resonator, 299
uartz, 98
  powdered –, 25

R

Race-track,
  magnet, 315
  orbit, 298
Radar, 426 ff.
  antenna, 362
  beacon, 432
  identification of place by, 432
  identification of vehicle by, 432
  network of ground-based – stations, 436
  reflection from meteor trails, 341
  screen, 432
  system, 432
  target, 441
  trace, 440
  use for collision prevention, 432
Radial oscillation, 290
Radial system for direction finding, 436, 444, 446
adiation,
  antenna, 369, 370
  black body –, 369 ff.
    law, 370
    from sun, 376
  incident noise –, 371 ff.
  loss in betatron, 295
  polarization of extraterrestrial, 378
  resistance, 369, 370
  solar, 364
  solar ionization –, 328
  thermal – from reflecting earth, 373
Radio,
  aids to navigation, 427, 448
  blackouts, 364
  frequency,
    measurement of – phase, 442
    radiation from sun, 376
    refractive index of air, 395
    variation of – phase, 431
  noise, 338, 347 ff., 380
    atmospheric – from thunderstorms, 388
    cosmic – (see under "Cosmic radio noise")
    diurnal variation, 340
    galactic, 348
    level, 387
    man-made –, 387
    natural, 388
  propagation prediction, basic, 387
  range, 429, 435
    multiple, 431
  receiver, 427
  signal,
    bandwidth, 437
    frequency, 437
    polarization, 428, 430
    pulse length, 440
  transmission, effect of, 450
  transmitter, 427
  wave,
    absorption of, 333
    attenuation, 397
    propagation, 317
    velocity of propagation of, 428, 450
Radioactive
  barium, use of, 22
  half lives, 265
  isotopes, determination of mass, 258
Radius,
  curvature of ion path, 221
  effective earth –, 401
Ratio bracket, 258
Ray,
  cathode – (see under "Cathode ray")
  extraordinary –, 325
  ordinary –, 325
  theory, 324
Rayleigh,
  -Jean's law, 375, 376, 378
  distribution, 406
Receiver,
  amplifier, tuned radio frequency, 349
  channels, 429

  noise of, 358, 419
    wide-band –, 440
  radio, 427
  sensitivity, 358
Receiving,
  antenna, 348, 390
  radio signal, 428
Reception,
  FM broadcast –, 369, 390
  noise, 348
Reckoning, dead, 426
Recombination coefficient, 324
Recorder, multifrequency, 342
Rectification,
  effect, 47
  at interface barrier of oxide cathodes, 47
Rectifier,
  copper oxide type, 33
  selenium, 34
Reflected primaries, 71, 85
Reflecting layer, 317
  height, 318
Reflection,
  atmospheric boundary layers, 402
  coefficient, 84, 105
    at atmosphere, 405
  from meteor trails, 340
  slow electrons, 84
  sporadic E –, 332
Reflex klystron, 81
Refraction,
  air, 395
  ionospheric, 356
Refractive index,
  air, 390
  discontinuities, 431
  distribution, 402
Refractive modulus, 397
Relativistic equation of motion, 282
Relativistic mass increase, 274
Resolving power of mass spectrograph, 241, 243
Resonance accelerator, 271 ff.
Resonator, quarter-wave cavity –, 299
Retarding field,
  emission in, 35 ff.
  method, 105
Richardson plot, 30, 31
Rods of the retina, 147
Rutherford scattering, 92

S

Sagittarius, constellation, 348, 349, 364, 374
Sample handling, 232
Saturation current, 30, 33
Scaling process, 195
Scanning, 133
  beam, high velocity, 154
  electron microscope, 168
  process, 132
Scattering,
  elastically reflected primaries, 85
  free electron –, 92
  Rutherford –, 92
Scene brightness (television), 135 ff.
Schottky,
  barrier, 34
  effect, 38
  emission, 47, 49
  region, 38, 39
Schuman and Ilford Q plates, 250
Screen,
  cathode ray –, 23
  radar, 432
Secondaries (secondary electrons), 65 ff., 152
  absorption of, 71, 76, 77
    before emission, 91
    rate of, 87, 88
  angular distribution, 79
  energy,
    data on, 83
    distribution, 96
    of emission of, 67
  production of, 70
    maximum depth of, 70
    by high speed primaries, 85
  properties of, 79
  range of primaries and, 87
  time of migration, 80
  transfer of energy to, 91
  velocity distribution of, 81, 82
    from insulators, 104
Secondary emission, 58, 65 ff.
  coefficient, 58, 71
  from insulators, 97
  measurement of,
    for insulating targets, 110
    for metallic targets, 93
  theory of –,
    insulators, 109
    metals, 90 ff.
  yield, 67 ff., 98, 99
    effect of adsorbed gas, 78
    effect of angle of incidence, 76
    effect of conductivity, 100
    effect of crystal structure, 73, 75
    effect of mechanical condition of surface, 77
    effect of primary current, 77

    effect of temperature, 75, 100
    effect of work function, 72
    of metals, 68
    saturation of, 99
Seebeck emf, 50
Selector, velocity, 242
Selenium rectifier, 34
Semiconductors,
  chemical potential, 29
  excess impurity –, 6
  Hall coefficient, 20
  impurity – of N type, 7, 9
  impurity type of, 6
  intrinsic –, 7
  luminescence in, 23
  modern theory, 5 ff.
  N-type –, 6, 7, 9
  phosphorescence in, 23
  P-type –, 6
Semi-infinite coplanar sheets, 183, 184
Series shift bracket, 257
Service range, 415
Servomechanism, use of, 429
Shielding problem, 305
Shimming of magnetic field, 275
Shore effect, 428
Shot,
  effect, 80
  noise, 142
Signal-to-noise ratio, 161, 163, 368, 369, 388
  threshold –, 138, 143
    computation of, 145
    for eye, 146
Silicated interface, 26
Skin depth, effect of, 3, 10
Sky-wave,
  component, 428
  curve, 450
  fluctuation, 449
  interference, 439
  pattern, 443
  system, 436
  transmission, 428, 433 ff.
    errors in, 443
Slip frequency, 297
Sloping layer, 428
Smallest resolved angle, 136
Smooth earth theory, 392
Sodium chloride (NaCl), 98 ff.
Solar,
  activity, 326, 384
    effect of, 326
    index of, 330
  ionization radiation, 328
  ionizing agent, 322
  noise, 339, 350
  observations, 331
  radiation, 364
Solid analysis, 263
Sonne system, 431
Source,
  electron bombardment –, 231
  gaseous discharge type, 228
  hot anode –, 228, 229
  ion –, 228
  spark –, 230
Space charge,
  barrier, 34
  effect, 36
  repulsion, 107
Spark,
  discharge, 230
  hot – ion source, 230
Sparking, 55
  phenomena, 47
Spectrograph, mass (see under "Spectrometer")
Spectrograph, optical, 263
Spectrographic measurements, mass, 254, 255
Spectrometer, mass, 168, 199 ff., 219 ff.
  aberration, second order, 207
  beta-ray –, 168
  commercially available –, 265
  crossed field –, 205, 208
  focusing, 208
    directional, 200
    perfect –, 211
    velocity, 240
  leak detection –, 264
  resolving power, 241, 243
  use of, 253
Spectroscopy, mass, 219 ff.
Spot distortion, 169, 183 ff.
  correction of, 195 ff.
  reduction of, 196
Spray discharge, 117
Stable orbit, 280
Standard atmospheric field, 391
Standard navigation system, 435
Star map, world, 352
Static methods, 111
Sticking potential, 97
Storage time, 150
  of eye, 146
Stratovision, 414
Strontium, 4, 25
  oxide, 5, 9, 23, 42, 98, 101
Sun,
  black body radiation, 376
  eruptions, 366, 376, 378, 379

- radio frequency radiation, 376
- zenith angle, 324

Sunspot,
- activity, 376, 378, 385
- cycle, 326, 385, 420
- eruptions, 366, 376, 378, 379
- number, 327, 383

Super XX film, 149, 158

Suppressor, noise, 388

Surface,
- wave, 391
- work function, 8

Synchro-cyclotron, 289, 300 ff.

Synchrotron, 289 ff., 394 ff.
- focusing, 292
- maximum energy, 299, 300
- oscillation, 295
- phase focusing, 290
- proton –, 312 ff.
- required frequency, 302
- start, 298

T

Target, 157
- beryllium, 74
- filament form –, 95
- insulating –, 152
- oxidized, 116
- photo-sensitive, 152
- probe –, 277
- television pickup tubes, 133 ff.
- thick –, 99

Teleran system, 435, 436

Television, 131 ff.
- amplifier, noise in, 155
- broadcast stations, 421
- color –, 391
- noise in – circuit, 162
- pickup tubes, 131 ff.
  - absolute performance scale, 139
  - curve, 150
  - design, 135
  - lens of, 137
  - noise, 145
  - non-storage type, 150
  - performance, 133
  - performance, ideal, 135, 141, 142
  - performance, measuring, 142, 143
  - storage type, 151
  - target, 133 ff.
- picture,
  - formation, 132, 133
  - noise visibility in, 160
- receiver, 132, 162
- spurious pattern in, 155
- transmitter, 132
- tubes, 167, 196, 199
  - spot distortion, 169

Terrain,
- effect of, 390
- irregularities, 391

Terrestrial agent, 322

Test pattern, 142

Thermal activation energy, 13

Thermionic,
- current density, 29
- emission, 2, 29 ff.
  - bombardment enhanced, 102
  - direct current –, 9, 21
- emitter, 101
- work function, 31, 35

Thermoelectric effect, 50

Thickness dependence, 59

Thin film,
- emission, 117
- method, 111
- phenomena, 114

Three-station system, 444

Threshold contrast, 136, 138, 143, 146

Time delay, 318

Time-difference method, 436

Timing errors, 434, 446

Torsion of optical axis, 201

Tracers problem,
- biological, 265
- chemical, 265

Transmission,
- coefficient, 32
- frequency, 390
- line, 362
- north-south –, 428
- path, 392
- by television to aircraft, 436

Transmitter, radio, 427

Transmitting antenna, 429
- FM –, 391
- height, 390

Trap for electrons, 8, 18 ff.

Triode method, 94, 102

Tropospheric,
- ducts, 392
- layer reflection, 419
- waves, 402
  - propagation, 405
  - random –, 408

Tubes,
- cathode ray, 167, 196
- television – (see under "Television pickup tubes")

Tungsten, 61, 68, 71
  filament, 36
Two-dimensional deflection field, 173

V

Van de Graaff machine, 271
Van der Waals' adsorption, 75
Vapor pressure, 11
Velocity,
  distribution method, 112
  filters, 242
  focusing, 222
    spectrograph of Aston, 240
  selection analyzers, 247
  selector, 242
Vibrating reed type of amplifier, 252
Viscous gas flow, 232
Visibility, noise, 160
Vision,
  human, 146
  range, 426
Visual process, 145
Visual sensation, 147, 148
Volatilization of solids, 237
Voltmeter, electron beam, 112

W

Wallman low noise circuit, 363
Wave,
  continuous –, system, 434
  front, orientation of, 428
  ground –,
    pattern, 443
    system, 436
    transmission, 428, 442
    transmission efficiency, 434
    transmission, phase measurement, 442, 450
  guide for linear accelerator, 309
  sky- –
    component, 428
    curve, 450
    fluctuation, 449
    interference, 439
    pattern, 443
    system, 436
    transmission, 428, 433 ff.
Whiddington law, 90
Whistle observation from meteor, 341
Wien filter, 247
Work function, 36, 69, 70
  lowering of, 72, 73
  measurement, 72, 73
  photoelectric, 34
  surface –, 8
  thermionic, 31, 35
  total –, 5
  tungsten, 229
World star map, 352

X

X-ray,
  critical potential, 72
  diffraction,
    analysis, 24
    pattern, 26
    technique, 5, 9, 11
  high voltage – machine, 270

Y

Yield, secondary emission (see under "Secondary emission"), 67 ff., 98, 99

Z

Zenith angle, sun's, 324
Zinc sulfides, secondary emission of, 98